

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-254883

(43)Date of publication of application : 25.09.1998

(51)Int.Cl.

G06F 17/30

(21)Application number : 09-054359

(71)Applicant : MITSUBISHI ELECTRIC CORP

(22)Date of filing : 10.03.1997

(72)Inventor : FUJII YOICHI

SUZUKI KATSUSHI

IMAMURA MAKOTO

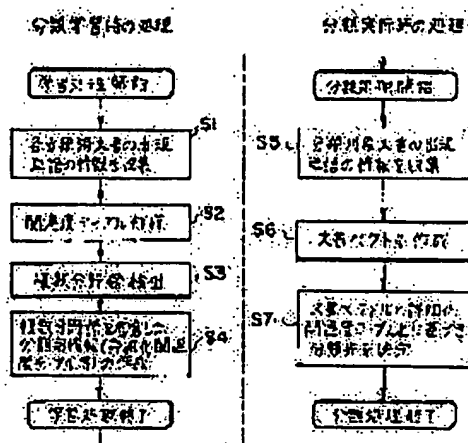
TAKAYAMA YASUHIRO

(54) AUTOMATIC DOCUMENT SORTING METHOD

(57)Abstract:

PROBLEM TO BE SOLVED: To provide an automatic document sorting method capable of precisely sorting even for detailed sorting.

SOLUTION: For learning, a word dividing/frequency extracting part collects information on appearing words from each sorted document (S1). A relation degree arithmetic part finds a relation degree between each word and each sort based on this information (S2). A plural-fields word processing part detects plural-fields word related with plural fields from this relation degree table (S3), divides each plural fields word by each strongly related field to regard as separated words and prepares information for sorting in a detailed relation degree table, etc., (S4). In sorting a document, a word dividing/frequency extracting processing part first collects information on the frequency, etc., of appearing words in the document (S5). A sorting destination deciding part 10 prepares a document vector expressing the tendency of appearing words in the sorting object document based on this information (S6) and decides the sorting destination of the document, based on this vector and the detailed relation degree table (S7).



HH

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP00/00203

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl⁷ G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Int.Cl⁷ G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1926-1996 Jitsuyo Shinan Toroku Koho 1996-2000
Kokai Jitsuyo Shinan Koho 1971-2000 Toroku Jitsuyo Shinan Koho 1994-2000

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	JP, 10-254883, A (Mitsubishi Electric Corporation), 25 September, 1998 (25.09.98), Claim 1 (Family: none)	1-5,12,14 6-8
Y A	Koichi Hashida, "GDA Versatile and Intelligent Content ware with Semantic Annotation", Journal of Japanese Society for Artificial Intelligence, Vol.13, No.4, 01 July, 1998 (01.07.98), pp.528-535	1-5,12,14 6-11,13,15
A	Katashi Nagao, Koiti Hasida, "Automatic Text summarization Based on the Global document Annotation", 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Proceedings of the Conference Volume II, (10.08.98), pp.917-921	1-15

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not
considered to be of particular relevance
"E" earlier document but published on or after the international filing
date
"L" document which may throw doubts on priority claim(s) or which is
cited to establish the publication date of another citation or other
special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other
means
"P" document published prior to the international filing date but later
than the priority date claimed

"T" later document published after the international filing date or
priority date and not in conflict with the application but cited to
understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be
considered novel or cannot be considered to involve an inventive
step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be
considered to involve an inventive step when the document is
combined with one or more other such documents, such
combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search
30 March, 2000 (30.03.00)

Date of mailing of the international search report
11 April, 2000 (11.04.00)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.



PATENT ABSTRACTS OF JAPAN

(11) Publication number: **10254883 A**(43) Date of publication of application: **25 . 09 . 98**

(51) Int. Cl

G06F 17/30(21) Application number: **09054359**(22) Date of filing: **10 . 03 . 97**(71) Applicant: **MITSUBISHI ELECTRIC CORP**(72) Inventor: **FUJII YOICHI
SUZUKI KATSUSHI
IMAMURA MAKOTO
TAKAYAMA YASUHIRO**(54) **AUTOMATIC DOCUMENT SORTING METHOD**

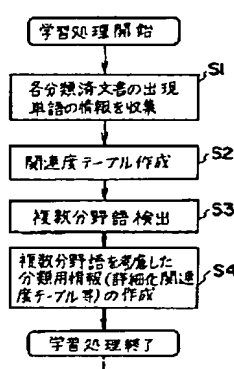
COPYRIGHT: (C)1998,JPO

(57) Abstract:

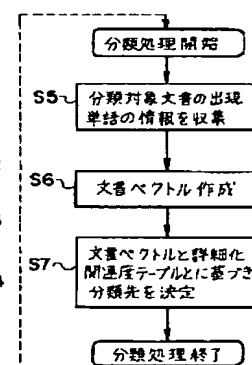
PROBLEM TO BE SOLVED: To provide an automatic document sorting method capable of precisely sorting even for detailed sorting.

SOLUTION: For learning, a word dividing/frequency extracting part collects information on appearing words from each sorted document (S1). A relation degree arithmetic part finds a relation degree between each word and each sort based on this information (S2). A plural-fields word processing part detects plural-fields word related with plural fields from this relation degree table (S3), divides each plural fields word by each strongly related field to regard as separated words and prepares information for sorting in a detailed relation degree table, etc., (S4). In sorting a document, a word dividing/frequency extracting part first collects information on the frequency, etc., of appearing words in the document (S5). A sorting destination deciding part 10 prepares a document vector expressing the tendency of appearing words in the sorting object document based on this information (S6) and decides the sorting destination of the document, based on this vector and the detailed relation degree table (S7).

分類学習時の処理



分類実行時の処理



(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-254883

(43) 公開日 平成10年(1998) 9月25日

(51) Int.Cl.⁶

G 0 6 F 17/30

識別記号

F I

G 0 6 F 15/401

3 1 0 D

審査請求 未請求 請求項の数 4 O L (全 27 頁)

(21) 出願番号 特願平9-54359

(22) 出願日 平成9年(1997) 3月10日

(71) 出願人 000006013

三菱電機株式会社

東京都千代田区丸の内二丁目2番3号

(72) 発明者 藤井 洋一

東京都千代田区丸の内二丁目2番3号 三

菱電機株式会社内

(72) 発明者 鈴木 克志

東京都千代田区丸の内二丁目2番3号 三

菱電機株式会社内

(72) 発明者 今村 誠

東京都千代田区丸の内二丁目2番3号 三

菱電機株式会社内

(74) 代理人 弁理士 吉田 研二 (外2名)

最終頁に続く

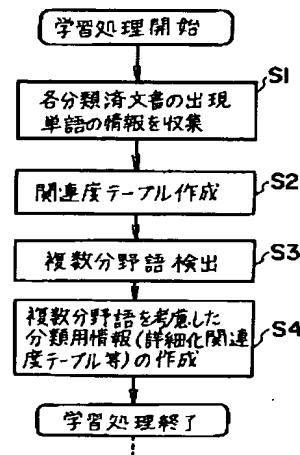
(54) 【発明の名称】 文書自動分類方法

(57) 【要約】

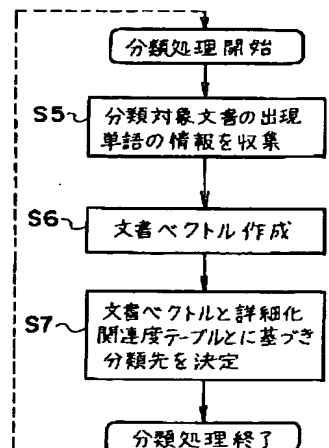
【課題】 細かい分類に対しても精度のよい分類を行うことができる文書自動分類方法を提供する。

【解決手段】 学習時には、単語分割／頻度抽出部は各分類済み文書から出現単語の情報を収集する (S1)。この情報に基づき関連度演算部が各単語と各分類との関連度を求め、関連度テーブルを作成する (S2)。複数分野語処理部は、この関連度テーブルから複数の分野に対して関連の強い複数分野語を検出し (S3)、各複数分野語を関連の強い各分野ごとに分割して別々の単語とみなして、詳細化関連度テーブルなどの分類用情報を作成する (S4)。文書を分類する際には、まず単語分割／頻度抽出処理部3が、当該文書の出現単語の頻度等の情報を収集する (S5)。分類先決定部10は、この情報に基づき当該分類対象文書の出現単語の傾向を表す文書ベクトルを作成し (S6)、このベクトルと詳細化関連度テーブルとに基づき当該文書の分類先を決定する (S7)。

分類学習時の処理



分類実行時の処理



【特許請求の範囲】

【請求項1】 分類済みの各文書に出現する各単語の頻度集計結果に基づき各単語と各分野との関連度を登録した関連度テーブルを作成し、この関連度テーブルから、閾値より高い関連度を有する強関連分野が複数存在する複数分野単語を求め、前記関連度テーブルにおける複数分野語についての欄を、当該複数分野語とこれに対応する強関連分野との組合せごとに複数の欄に分割して詳細化関連度テーブルを作成する分類学習ステップと、分類対象の文書に出現する単語の頻度を集計し、この結果得られた頻度情報を前記複数分野語の情報によって詳細化し、この詳細化された頻度情報と前記詳細化関連度テーブルとに基づき当該文書の分類先の分野を決定する分類実行ステップと、を含むことを特徴とする文書自動分類方法。

【請求項2】 単語の前記強関連分野の判定基準となる前記閾値は、前記関連度テーブルにおける当該単語の各分野に対する関連度の中の最大値に基づき定められることを特徴とする請求項1記載の文書自動分類方法。

【請求項3】 前記分類学習ステップでは、各複数分野語について、当該複数分野語の各強関連分野ごとに、その強関連分野に属する分類済み文書において当該複数分野語と共に出現した単語の傾向を表す共起ベクトルを生成し、

前記分類実行ステップでは、各複数分野語ごとに、この分類対象文書において当該複数分野語と共に出現した単語の傾向を示す文書共起ベクトルを生成し、この文書共起ベクトルと前記各共起ベクトルとの類似性に基づき、分類対象文書から得られた前記頻度情報を詳細化することを特徴とする請求項1又は2記載の文書自動分類方法。

【請求項4】 各単語の概念的な階層関係を記述したシソーラスを利用して、前記各共起単語の上位概念の情報を反映した共起ベクトル及び文書共起ベクトルを生成することを特徴とする請求項3記載の文書自動分類方法。

【発明の詳細な説明】**【0001】**

【発明の属する技術分野】この発明は、文書を自動分類する文書自動分類方法に関し、特に文書に出現する各単語の頻度の情報に基づき文書の分類先を決定する文書自動分類方法に関する。

【0002】

【従来の技術】文書自動分類の方式のなかの有力なもの一つに、分類先の分野が既知の文書から出現単語の頻度統計をとって各分野に固有のキーワードや各分野の単語出現傾向などを学習し、これらを分類基準として用いて文書を分類するという方式がある。

【0003】このような方式の文書自動分類装置の一つに、例えば特開平6-348755号公報に示される装置がある。この装置では、分類済みの文書群から各分野に固有の単語（キーワード）を抽出し、分類対象文書の

中のこれらキーワードの頻度に基づき当該文書の分類先を決定している。

【0004】図32は、この従来装置の構成図である。以下、分野のキーワードを学習する学習処理と、分類対象文書を分類する分類処理とに分けて、この従来装置の構成及び機能について説明する。

【0005】まず、学習処理では、分類済み文書データ101から全ての文書が取り出され、文書データ単語分割部103で単語分割がされ、この分割結果の情報が分類済み文書単語分割テーブル104に格納される。分類用辞書作成部106は、分類済み文書単語分割テーブル104に格納された単語分割テーブルの情報に基づき、特定の分野のみに現れる単語を当該分類のキーワードとして抽出し、これらキーワードの（見出し文字列、品詞、分野名）の組を分類用辞書107に格納する。このようにして分類用辞書107が完成すると、学習処理は終了する。図33は、この学習処理の結果得られた分類用辞書107の内容例である。この例は、例えば単語「自然語」は、分類名「言語処理」のキーワードであり、学習した分類済み文書群の「題名」に0回、「要旨」に1回、「目的」に1回、その他合計して17回出現したことを示している。

【0006】次に分類処理では、分類対象文書データ102から取り出された分類対象文書に対して、学習時と同様に文書データ単語分割部103で単語分割が行われ、この結果得られた単語分割データが分類対象文書単語分割テーブル105に格納される。文書分類部110は、分類対象文書単語分割テーブル105に格納された分類対象文書の単語分割データと、分類用辞書107の（見出し文字列、品詞、分野名）との間で、同じ単語を含むものの一致回数を各「分野名」毎に集計し、一致回数の最も多い「分野名」を当該分類対象文書の分類先の最優先候補として文書分類結果111に格納する。

【0007】この文書分類結果111は、分類結果確認部112に表示される。ユーザは、この分類結果が正しいか否かを判断し、もしこの分類結果が間違っただと判断した場合には、誤分類の原因となった単語を判定して入力する。すると、分類用辞書学習部113が、この誤分類原因単語のデータ（見出し文字列、品詞、分野名）を分類用辞書107から削除する。この装置では、このような構成により、学習に用いた分類済み文書群の偏りによるキーワード選択の不備を補正できるようにしている。

【0008】また、別の従来装置として、特開平7-114572号公報に示される装置がある。この装置では、文書中の単語の共起関係に基づき単語の特徴ベクトルを生成し、その単語特徴ベクトルから文書の特徴ベクトルを作成し、文書の特徴ベクトル同士の間類似度を利用して文書を分類する。

【0009】図34は、この従来装置の構成図である。

以下、学習時の処理と分類時の処理とに分けて、この従来装置の構成及び機能について説明する。

【0010】まず学習処理では、まず文書解析部122が、文書記憶部121内の分類先が既に定まっている学習用文書を取り出し、この文書を解析して出現単語を抽出する。文書解析部122で全ての学習用文書から単語の抽出が完了すると、単語ベクトル生成部123は、各単語について、その単語と同一文書中に現れる単語

(「共起単語」と呼ぶ)を集計し、共起単語を基底としその頻度を成分値とする単語ベクトルを生成する。この単語ベクトルは、単語についての共起単語の傾向を表すベクトルである。得られた単語ベクトルは、単語ベクトル記憶部124に記憶される。なお、ここで、文書解析部122、単語ベクトル生成部123にて処理対象となる単語は、単語ベクトル生成用辞書129に登録されている単語に限定される。

【0011】図35は、このようにして得られた単語ベクトルの例を示すものである。単語ベクトルの各成分の上に記されている単語が、それら各成分の基底である。図35では、例えば単語「アメリカ」という単語に対しては、単語「政府」、「先進」、「主要」、「国」、「ココム」が同一文書中に現れた(共起した)ことがあり、共起回数は各1回ずつであったことを示している。

【0012】対象となる全単語に対する特徴ベクトルの単語ベクトル記憶部124への記憶が完了すると、文書ベクトル生成部125は、文書解析部122で一つの文書から抽出された各単語について、対応する単語ベクトルを当該単語の出現頻度で重み付けして加算することにより、当該文書の特徴を表す文書ベクトルを生成する。例えば、単語「アメリカ」、「兵器」がそれぞれ1回ずつ現れた文書の文書ベクトルは、図36に示すような形となる。求められた文書ベクトルは、文書ベクトル記憶部126に記憶される。

【0013】次に、文書ベクトル記憶部126に記憶された各学習用文書の文書ベクトルに基づき、各分野の特徴を示す代表ベクトルを生成する。ある分野の代表ベクトルは、同一分類に含まれる各学習用文書の文書ベクトルを加算することにより求められる。求められた代表ベクトルは、後述の分類処理において類似度を計算するときにベクトルの大きさが影響しないよう長さ1のベクトルに正規化される。以上で学習処理が終了する。

【0014】次に分類処理について説明する。まず、文書解析部122は、指示された分類対象文書を文書記憶部121から取り出し、文書解析を行って出現単語を抽出する。次に、文書ベクトル生成部125は、各出現単語に対応する単語ベクトルを出現頻度で重み付けして加算し、当該分類対象文書の文書ベクトルを生成して文書ベクトル記憶部126に記憶する。そして、分類部127が、この分類対象文書の文書ベクトルに最も類似した代表ベクトルを例えばベクトルの内積演算によって求

め、この最類似代表ベクトルに対応する分野に当該分類対象文書を分類する。この分類結果は、分類結果記憶部128に記憶される。

【0015】また、さらに別の従来装置として、『意味属性の学習結果にもとづく文書自動分類方式』(河合、情報処理学会論文誌Vol. 33, No. 9, pp. 1114-1122)に示される装置がある。

【0016】以下、図37を参照して、この従来装置について説明する。

【0017】まず、学習段階では、名詞抽出処理部133は、学習用文書131を取り出し、単語辞書144の情報をを用いて、当該学習用文書から名詞を抽出する。次に、意味属性抽出処理部136は、学習用文書名詞抽出結果134に格納された各名詞の意味属性を、シソーラス143から抽出する。シソーラス143に格納されている名詞と意味属性の関係の例を図38に示す。図38には、例えば「醤油」という名詞が「調味料」という意味属性を有していることが示されている。各名詞の意味属性の抽出結果は、学習用文書意味属性抽出結果137に格納される。次に、分類用辞書作成部139は、学習用文書名詞抽出結果134における各名詞の頻度を分野毎に集計するとともに、学習用文書意味属性抽出結果137における各意味属性の頻度を分野毎に集計する。図39は、各意味属性の各分野毎の頻度の集計結果のテーブルの一例であり、例えば「人工物」という意味属性を持つ単語が「運輸通信」の分野に属する文書に12回出現したことを示している。単語の集計結果についても同様のテーブルが作成される。分類用辞書作成部139は、これら各集計結果のテーブルに対し、統計学におけるカイ2乗検定の考え方を応用した計算式を適用することにより、各名詞と各分野と関連度合いを表したテーブル、及び各意味属性と各分野との関連度合いを表したテーブルを作成する。作成された各テーブルは、分類用辞書140に格納される。以上で学習処理が終了する。

【0018】次に分類時の処理について説明する。まず、名詞抽出処理部133は、分類用文書132から分類対象に指定された文書を取り出し、単語辞書144の情報をを用いて、当該文書から名詞を抽出し、分類用文書名詞抽出結果135に格納する。次に、意味属性抽出処理部136は、抽出された各名詞の意味属性をシソーラス143から抽出し、分類用文書意味属性抽出結果138に格納する。そして、文書分類部141は、まず分類用文書名詞抽出結果135と、分類用辞書140の単語と分野の関連度合いのテーブルとに基づき、単語の出現頻度からみた当該文書の各分野への関連度合いを計算する。また、文書分類部141は、分類用文書意味属性抽出結果138と、分類用辞書140の意味属性と分野との関連度合いのテーブルとに基づき、意味属性の頻度から見た当該文書の各分野への関連度合いを計算する。そして、文書分類部141は、両計算結果を所定の比率で

加算することにより、各分野ごとに、当該分類対象文書と当該分類の関連度合いを求める。そして、例えばこの関連度合いの値の最も大きい分野が当該分類対象文書の分類先に選ばれ、文書分類結果142に格納される。以上で自動分類処理が終了する。

【0019】

【発明が解決しようとする課題】以上に説明した従来の各文書自動分類装置は、いずれも、分類済み文書において各分野に特徴的に出現する単語を学習し、この学習結果を分類基準として文書を分類する点では一致する。例えば、特開平7-114572号公報に示された技術では、ある分野に特徴的に現れる単語は、当該分野の代表ベクトルにおいて大きな値の成分となるので、類似度の値に大きな影響を与え、分類先の決定に大きな影響を与える。また、河合の論文に示された技術でも、ある分野に特徴的に現れる単語は、その分野との関連度合いの値が大きくなるので、分類先を大きく左右する。

【0020】このような手法は、例えば<政治>、<経済>などのようにある程度関連が強い単語が共通して出現する分野同士の間では、分類を誤る可能性が高い。例えば、「首相」という単語は、<政治>の分野の文書（例えば新聞記事）によく出現する（すなわち特徴的な単語である）が、<経済>の分野の文書にもある程度出現する。ここで、<経済>分野の文書にたまたま「首相」という単語が多く含まれていると、その文書は「首相」という単語の影響で<政治>分野に誤分類されてしまう可能性が高い。

【0021】このように、従来の文書自動分類技術では、複数の分野に対してそれぞれある程度強い関連を有する単語が分類対象の文書内に数多く現れると、その文書はその単語に対する関連が最も強い分野に分類されやすく、このため誤分類が生じる可能性が高かった。

【0022】このような傾向は、例えば、<政治>、<科学>、<スポーツ>などのように、出現する単語の傾向の相違が大きい分野同士の間での大まかな分類では致命的な問題にはならないかも知れない。しかしながら、分類を細かくしようとすると、類似する分野が増えてくるので、複数の分野が共通の単語にある程度以上の関連を有するような場合がどうしても増えてくる。このため、従来の文書分類技術は、分類が細くなると誤分類の可能性が増し、分類精度が劣化するという問題があった。

【0023】本発明は、このような問題点を解決するためになされたものであり、類似する分野間での分類の精度を向上させることにより、細かい分類に対しても精度のよい分類を行うことができる文書自動分類方法を提供することを特徴とする。

【0024】

【課題を解決するための手段】この発明は、上記課題を解決するためになされたものであり、分類済みの各文書

に出現する各単語の頻度集計結果に基づき各単語と各分野との関連度を登録した関連度テーブルを作成し、この関連度テーブルから、閾値より高い関連度を有する強関連分野が複数存在する複数分野単語を求め、前記関連度テーブルにおける複数分野語についての欄を、当該複数分野語とこれに対応する強関連分野との組合せごとに複数の欄に分割して詳細化関連度テーブルを作成する分類学習ステップと、分類対象の文書に出現する単語の頻度を集計し、この結果得られた頻度情報を前記複数分野語の情報によって詳細化し、この詳細化された頻度情報と前記詳細化関連度テーブルとに基づき当該文書の分類先の分野を決定する分類実行ステップとを含むものである。

【0025】また、単語の強関連分野の判定基準となる閾値を、関連度テーブルにおける当該単語の各分野に対する関連度の中の最大値に基づき定めるものである。

【0026】また、分類学習ステップでは、各複数分野語について、当該複数分野語の各強関連分野ごとに、その強関連分野に属する分類済み文書において当該複数分野語と共起した単語の傾向を表す共起ベクトルを生成し、分類実行ステップでは、各複数分野語ごとに、この分類対象文書において当該複数分野語と共起した単語の傾向を示す文書共起ベクトルを生成し、この文書共起ベクトルと前記各共起ベクトルとの類似性に基づき、分類対象文書から得られた前記頻度情報を詳細化するものである。

【0027】また、各単語の概念的な階層関係を記述したシソーラスを利用して、各共起単語の上位概念の情報を反映した共起ベクトル及び文書共起ベクトルを生成するものである。

【0028】

【発明の実施の形態】

実施の形態1. 以下、この発明の実施の形態を図面を参照して説明する。

【0029】図1は、この発明に係る文書自動分類方法を実施するためのシステムの構成図である。図1において、分類済み文書記憶部1は、分類済み文書の文書データと、それら各分類済み文書の分類先分野が登録された分類先リストと、を記憶している。これら分類済み文書や分類先リストは、分類基準の学習のために用いられる。また、分類対象文書記憶部2は、自動分類の対象となる分類対象文書の文書データを記憶している。

【0030】単語分割/頻度抽出部3は、分類済み文書記憶部1又は分類対象文書記憶部2から供給される文書に対し例えば形態素解析などを行うことにより、その文書を単語に分割し、これら単語の頻度統計をとる。そして、単語分割/頻度抽出部3は、その文書内でのそれら各単語の出現位置などの情報を含んだ単語分割情報、及びその文書内での各単語の出現頻度を示す頻度情報を作成する。分類済み文書についての単語分割情報及び頻度

情報（単語分割／頻度情報 5 1）は、分類済み文書単語分割／頻度情報記憶部 4 に記憶される。一方、分類対象文書についての単語分割情報及び頻度情報（単語分割／頻度情報 5 8）は、分類対象文書単語分割／頻度情報記憶部 5 に記憶される。

【0031】関連度演算部 6 は、分類の基準を学習する際において、分類済み文書記憶部 1 に記憶された各分類済み文書の文書分類先テーブル 5 2 の情報と、分類済み文書単語分割／頻度情報記憶部 4 に記憶された頻度情報とに基づき、各分野ごとに出現単語の頻度を集計し、この頻度集計の結果に基づき各単語と各分野との関連度を計算する。頻度集計結果及び各単語と各分野との関連度は、それぞれ頻度集計テーブル 5 3 及び関連度テーブル 5 4 として関連度情報記憶部 7 に記憶される。

【0032】複数分野語処理部 8 は、分類学習時には、関連度情報記憶部 7 の関連度テーブル 5 4 に基づき、関連度の高い分野（「強関連分野」と呼ぶ）が複数存在する単語（「複数分野語」と呼ぶ）を検出し、複数分野語リスト 5 5 を作成する。また、複数分野語処理部 8 は、この複数分野語リスト 5 5 の情報を用い、分類済み文書単語分割／頻度情報記憶部 4 の頻度集計テーブル 5 3 を詳細化し、この詳細化された頻度集計テーブルに基づき詳細化関連度テーブル 5 6 を作成する。すなわち、ここでは、各複数分野語を（単語、強関連分野）の組合せごとに別々の単語と捉え直し、単なる単語と分野との関連度だけでなく、複数分野語については（単語、強関連分野）の組合せと各分野との関連度をも含んだ詳細化関連度テーブル 5 6 を作成する。また、複数分野語処理部 8 は、複数分野語リスト 5 5 と分類済み文書単語分割／頻度情報記憶部 4 の単語分割情報とに基づき、各複数分野語と例えば同一文書内や同一段落内に現れる単語（共起単語）を求め、これら共起単語の出現傾向を表す共起ベクトル 5 7 を（複数分野語、強関連分野）の各組合せごとについて作成する。

【0033】このようにして求められた複数分野語リスト 5 5、詳細化関連度テーブル 5 6 及び各共起ベクトル 5 7 は、分類用情報記憶部 9 に格納される。この分類用情報記憶部 9 に格納された情報が、文書の分類先を決定するための基準となる。

【0034】また、複数分野語処理部 8 は、文書分類時には、分類対象文書に含まれる各複数分野語について、例えば同一段落における共起単語を検出し、当該分類対象文書における共起単語の傾向を表す文書共起ベクトル 5 9 を作成する。

【0035】分類先決定部 10 は、分類対象文書の出現単語の頻度情報及び文書共起ベクトル 5 9 と分類用情報記憶部 9 に記憶された各共起ベクトル 5 7 とに基づき、当該文書の分類上の特徴を表す文書ベクトル 6 0 を作成する。そして、分類先決定部 10 は、この文書ベクトル 6 0 と分類用情報記憶部 9 に記憶された詳細化関連度テ

ーブル 5 6 とに基づき、当該文書と各分野との関連度を計算し、この関連度の値に基づき当該文書の分類先の分野を決定する。この結果得られた文書分類結果 6 1 は、文書分類結果記憶部 11 に格納される。

【0036】次に、図 2 を参照して、この実施の形態のシステムの処理手順の全体的な流れを説明する。図 2 に示すように、このシステムの処理手順は、分類の基準となる情報を学習する分類学習時の処理と、与えられた分類対象文書を分類する分類実行時の処理とに分かれる。

【0037】分類学習時においては、まず S 1 にて、単語分割／頻度抽出部 3 により、分類済み文書記憶部 1 の各分類済み文書について、出現単語の情報（すなわち単語分割／頻度情報 5 1）を収集する。S 2 では、これらの情報を用いて、関連度演算部 6 により各単語と各分類との関連度を保持する関連度テーブル 5 4 を作成する。S 3 では、複数分野語処理部 8 が、この関連度テーブル 5 4 から複数分野語を検出し、複数分野語リスト 5 5 を作成する。そして、S 4 では、この複数分野語リスト 5 5 に基づき、複数分野語処理部 8 が、分類の基準となる情報として、詳細化関連度テーブル 5 6 や共起ベクトル 5 7 など、複数分野語を考慮した分類用情報を作成する。

【0038】そして、個々の未分類の文書の分類を実行する際には、まず S 5 にて、単語分割／頻度抽出処理部 3 が、分類対象文書の出現単語の情報（すなわち単語分割／頻度情報 5 8）を収集するとともに、複数分野語処理部 8 が、単語分割／頻度情報 5 8 に基づき各複数分野語について文書共起ベクトル 5 9 を作成する。次に、S 6 では、分類先決定部 10 が、単語分割／頻度情報 5 8、文書共起ベクトル 5 9、及び分類用情報記憶部 9 内の各共起ベクトル 5 7 を用いて、当該分類対象文書の分類上の特徴を表す文書ベクトル 6 0 を作成する。そして、S 7 では、分類先決定部 10 が、文書ベクトル 6 0 と詳細化関連度テーブル 5 6 とに基づき、当該分類対象文書の分類先の分野を決定する。

【0039】以下、図 2 の各ステップの処理を更に詳細に説明する。

【0040】まず図 3 は、単語分割／頻度抽出部 3 の処理手順の具体例を示すフローチャートである。図 2 の S 1（分類済み文書からの出現単語情報の収集処理）では、単語分割／頻度抽出部 3 が、この図 3 の処理手順に従って、各分類済み文書についての単語分割情報及び頻度情報を収集する。以下、図 3 の処理手順を詳説する。

【0041】単語分割／頻度抽出部 3 は、分類済み文書記憶部 1 から分類済み文書を読み込むと、まず S 11 において当該文書から 1 段落を切り出し、この段落の情報を保持する。次に、S 12 にて、切り出した段落から 1 文を切り出し、この文の情報を保持する。次に、S 13 にて、この文に対して形態素解析を行うことにより、この文から順次単語を切り出す。S 14 では、分割された

単語の文書中の位置と品詞の情報を求め、これら情報を分類済み文書単語分割／頻度情報記憶部4内の当該分類済み文書の単語分割情報に登録する。そして、S15では、分類済み文書単語分割／頻度情報記憶部4内の当該分類済み文書の頻度情報において、S13にて切り出された単語の頻度を1だけ増やし、頻度情報を更新する。なお、S14とS15は、いずれを先に行ってもよい。

【0042】S14及びS15の処理は、単語の切り出しが文の末尾に達するまで繰り返される。文の末尾に達すると、S12に戻って、段落から次の文を切り出す。また、文の切り出しが段落の末尾に達すると、S11に戻って、分類済み文書から次の段落を切り出す。このようにして、分類済み文書の末尾に達するまでS11、S12、S13、S14及びS15の処理が繰り返される。この結果、当該分類済み文書についての単語分割情報及び頻度情報が完成する。この図3の処理は分類済み文書記憶部1に格納された各分類済み文書ごとに繰り返される。

【0043】この図3の処理を、具体例を用いて説明する。

【0044】例えば、単語分割／頻度抽出部3に、分類済み文書として図9に示す文書が与えられたとする。この場合、単語分割／頻度抽出部3は、まずS11にて最初の段落70を切り出し、分類済み文書単語分割／頻度情報記憶部4に作成した当該文書の単語分割情報に、段落の先頭を示す情報を登録する。

【0045】図12は、図9の文書から作成された単語分割情報の一例を示す図である。この例では、情報タイプとしてP、W、Sの3つのコードが設けられている。コードPは段落の先頭を示すコードであり、あるPとその次のPとに挟まれた部分が、一つの段落についての情報となる。また、コードSは文の先頭を示すコードであり、あるSとその次のSとに挟まれた部分が、一つの文についての情報である。そして、コードWは単語を示すコードであり、情報タイプがWの欄には、当該単語の文書冒頭からの位置（バイト単位で示される）と、当該単語を表す文字列と、当該単語の品詞が登録される。図12は、例えば、図9の文書の最初の段落の最初の文において、文書冒頭から3バイト目の位置に、「内閣」という名詞が出現していることを示している。

【0046】したがって、S11で段落が切り出されると、単語分割情報には情報タイプにコードPが登録される。

【0047】次に、単語分割／頻度抽出部3は、S12にて段落70から1文を切り出し、単語分割情報に文の先頭を表すコードSを情報タイプとして登録する。図9では、段落70は1文のみしか含まないので、この段落70（＝1文）について、S13以下の処理が行われる。すなわち、S13では、従来公知の手法である形態素解析を使ってこの1文を解析し、文中に含まれる単語

に分割する。S14では、文頭から順番にまず単語「内閣」を取り出す。そして、S14では、切り出した単語「内閣」について、文書冒頭からの位置と品詞を求める。この結果、単語分割情報には、図12に示すように、情報タイプ「W」、単語位置「3」、単語「内閣」及び品詞「名詞」が登録される。

【0048】そして、S15では、単語分割／頻度抽出部3は、分類済み文書単語分割／頻度情報記憶部4に作成した当該文書の頻度情報に対し、S13で分割した単語「内閣」を反映させる。すなわち、単語「内閣」は初出なので、頻度情報に単語「内閣」の欄を作成し、その頻度を1と設定する。なお、初出でない単語の場合は、頻度情報における当該単語の欄の頻度に1を加える。

【0049】このようにして単語「内閣」についての処理が終わると、段落70の文から次の単語「支持率」が取り出され、S14及びS15でこの単語の情報が単語分割情報及び頻度情報に反映される。

【0050】このようにして、S13で分割した文末までの各単語についてS14及びS15の処理が終了すると、S12に戻って段落70から次の文書を切り出そうとするが、この例では段落70は1文しか含まないので、S11に戻り、分類済み文書から次の段落を切り出し、以上の処理を繰り返す。そして、このような処理を、当該分類済み文書の末尾まで繰り返すことにより、図11に示す頻度情報と、図12に示す単語分割情報が得られる。

【0051】図11に示す頻度情報では、各単語の頻度についての情報が保持されている。例えば、図11は、図9の分類済み文書に、「首相」という名詞が3回出現し、「会見」という「サ変名詞」（語尾に「する」を付加することにより動詞として用いることができる名詞）が1回出現したことを示している。なお、以下では、必要に応じ、単語を「単語名：品詞」という形式で表現する。

【0052】このようにして、分類済み文書記憶部1に格納された全ての分類済み文書に対して図3に示す処理を行うことにより、各分類済み文書の単語分割情報及び頻度情報が得られる。これらの情報は、分類済み文書単語分割／頻度情報記憶部4に記憶される。

【0053】次に、図2のS2（関連度テーブル作成）の処理について、図4を参照して説明する。前述したように、このS2の処理は、関連度演算部6によって実行される。図4は、関連度演算部6が実行する処理の手順を示したものである。

【0054】まず、関連度演算部6は、S21にて、分類済み文書記憶部1の文書分類先テーブル52から分類済み文書を1つ選択し、この文書の分類先を取得すると共に、分類済み文書単語分割／頻度情報記憶部4からこの文書の頻度情報（例えば図11参照）を取り出す。例えば、図10は文書分類先テーブル52のデータ内容の

一例を示している。図10に示すように、文書分類先テーブル52には、分類済み文書記憶部1に記憶された各分類済み文書ごとに、その文書の文書名（例えばファイル名）と分類先の分野名とが登録されている。一方、分類済み文書単語分割／頻度情報記憶部4には、各分類済み文書の頻度情報及び単語分割情報が、その文書の文書名に対応づけて格納されている。関連度演算部6は、この文書分類先テーブル52のエントリ（すなわち分類済み文書）を先頭から順次選択し、選択した文書の頻度情報を文書名で検索する。

【0055】次に、S22では、関連度演算部6は、S21で取り出した頻度情報に基づき、頻度集計テーブル53における当該分類済み文書の分類先の各分野について、当該頻度情報に登録された各単語の頻度値をカウントアップする。頻度集計テーブル53は、例えば、各分野を横軸とし各単語を縦軸とするテーブルである。なお、分類済み文書の分類先が複数分野にわたる場合には、このカウントアップ処理においては、頻度情報における各単語の頻度値をその分類先の分野の数で割った値を頻度集計テーブル53の該当欄に加える。

【0056】以上のステップを具体例を用いて説明する。S21で例えば図9の分類済み文書が選択され、この文書の文書名が『11/04M/04-09』であったとする。この場合、関連度演算部6は、図10の分類先リストからこの文書に対応する分野<政治>（以下、分野名は<>で括って表現する）を検出すると共に、図11に示された頻度情報を分類済み文書単語分割／頻度情報記憶部4から取り出す。次に、S22では、関連度演算部6は、図11の頻度情報から、「会見：サ変名詞」の頻度値が1、「首相：名詞」の頻度値が3、などの各単語の頻度値を取得する。この例では、処理対象の文書の分類先は唯一<政治>のみであるので、それら各単語の頻度値は、そのまま頻度集計テーブル53の分野<政治>の列における各単語の頻度に加えられる。なお、処理対象が図10の文書『11/12M/09-08』である場合には、この文書の分類先分野は2つあるので、各単語の頻度を2で除したものが、頻度集計テーブル53のそれら2分野における各単語の頻度それぞれ加えられる。

【0057】すべての分類済み文書についてS21及びS22の処理が終了すると、頻度集計テーブル53が完成する。完成した頻度集計テーブル53のデータ内容の一例を図13に示す。図13は、例えば、分類済み文書群の中の<政治>分野に分類される文書において「首相：名詞」という単語が50回出現したことを示している。この頻度集計テーブル53は、関連度情報記憶部7に格納される。

【0058】頻度集計テーブル53が完成すると、次に関連度演算部6は、S23にて、このテーブルにおける各単語各分野についての頻度集計結果から、以下に示す

式(1)を用いて各単語各分野の理論頻度 M_{ij} を算出する。

【0059】

【数1】

$$M_{ij} = \frac{\sum_{i=1}^N F_{ij}}{\sum_{j=1}^L (\sum_{i=1}^N F_{ij})} \cdot \sum_{j=1}^L F_{ij} \quad (i=1, \dots, N, j=1, \dots, L)$$

10 ここで、Nは頻度集計テーブル53における分野の数、Lは同テーブルにおける単語の数を示し、iは各分野に付した続き番号、jは各単語に付した続き番号を示す。また、 F_{ij} は分野iの文書における単語jの出現頻度を示す。

【0060】式(1)においては、理論頻度 M_{ij} は、（単語jの総頻度が全単語の総頻度に占める割合）×（分野iにおける各出現単語の総数）の形で定義されている。すなわち、理論頻度 M_{ij} は、単語jが特定の分野に偏らず各分野に平均的に出現すると仮定した場合において、単語jが分野iに出現する期待される頻度を意味する。この理論頻度については、前述の河合の論文に説明されている。

【0061】このようにして分野iと単語jとの組合せごとに理論頻度 M_{ij} が求められると、次に関連度演算部6は、S24にて、この理論頻度の計算結果と、この元となった頻度集計テーブル53に基づき、分野iと単語jとの関連度 Y_{ij} を計算する。ここで、関連度 Y_{ij} は、カイ2乗検定を応用した以下の計算式を用いて計算する。

30 【0062】

【数2】

$$Y_{ij} = (F_{ij} - M_{ij}) \{ F_{ij} - M_{ij} / M_{ij} \} \quad (i=1, \dots, N, j=1, \dots, L)$$

関連度 Y_{ij} の計算結果は、関連度テーブル54の形で関連度情報記憶部7に格納される。図14に、図13の頻度集計テーブルから求められた関連度テーブル54のデータ内容を例示する。

【0063】式(1)及び式(2)を用いて求められた分野iと単語jとの関連度 Y_{ij} は、関連が強い（分野、単語）の組合せでは正の大きい値となり、逆に関連が弱い（分野、単語）の組合せについては負の大きな値となる。また、幾ら分野iに対して単語jの出現頻度が大きかったとしても、その単語jが全分野に平均して多く出現する場合は、分野iと単語jとの関連度 Y_{ij} は小さい値（0に近い値）となる。すなわち、ある分野iに特異的によく出現する単語jがあったとすると、両者の関連度 Y_{ij} は正の大きい値となる。また、単語jが分野iの文書にほとんど出現しなかったような場合には、両者の関連度 Y_{ij} は負の大きい値となる。なお、この関連度の考え方も前述の河合の論文に示されている。

【0064】関連度テーブル54が完成すると、次に複数分野語処理部8は、S3にて、このテーブルに基づき複数分野語を検出する。このS3の処理の詳細を図5を参照して説明する。

【0065】まず、複数分野語処理部8は、S31で関連度テーブル54に欄のある全単語について処理が完了したかを検査する。完了していない場合には、S32で関連度テーブル54から単語を1つ選択する。そして、複数分野語処理部8は、S33で、関連度テーブル54を参照し、この単語について関連の強い分野（すなわち強関連分野）を検出し、強関連分野が複数個あった場合にはS34に移る。S34では、複数分野語リスト55にこの単語についての欄を作成し、この単語に対する複数の強関連分野の分野名をこの欄に登録する。なお、S33の判定において、強関連分野が1つしかなかった場合には、この単語については何も行わず、S31に戻る。そして、関連度テーブル54に登録された全単語について以上の処理を繰り返すことにより、強関連分野が複数ある単語と、その単語に対応する強関連分野（複数）とが登録された複数分野語リスト55が完成する。

【0066】この処理手順において、ある分野が強関連分野であるか否かは、当該分野の関連度を閾値と比較することにより判定する。閾値より大きい関連度を持つ分野は強関連度分野と判定する。この判定のための閾値は、選択した単語についての各分野の関連度のうちの最大値に所定の係数（1以下の正数）を乗じた値を用いる（したがって、どの単語も最低一つは強関連分野を有する）。

【0067】すなわち、複数分野語処理部8は、選択した単語jについて、関連度テーブル54におけるj番目の行の各分野の関連度を取り出し、これら関連度に次式（3）を適用することにより、当該単語jの強関連度分野の数 T_j を算出する。

【0068】

【数3】

$$T_j = \# \{ Y_{ij} | Y_{ij} > \max_{i=1}^N (Y_{ij}) \cdot Y_{border} \} \quad (j=1, \dots, L)$$

ここで、演算子『#』は、後続の集合『{ }』に含まれる要素（ Y_{ij} ）の数を返す。また、 Y_{border} は、強関連分類の判定のための閾値を求める際の所定の係数（固定値）である。

【0069】この式（3）によれば、単語jの各分野に対する関連度の最大値に対して所定の割合となる値を閾値とし、関連度 Y_{ij} がその閾値より大きい分野（すなわち強関連分野）の数が T_j として求められる。この T_j の値が2以上の場合、単語jは複数分野語と判定され、複数分野語リスト55にその単語名と各強関連分野の分野名が登録される。

【0070】具体例を用いて説明する。例えば、図14の関連度テーブルが与えられ、係数 Y_{border} の値を0.

3にした場合、単語「首相」については、最大の関連度を持つ分野<政治>の関連度（66.7）にこの係数0.3を乗じたものが、この単語「首相」についての強関連分野判定の閾値となる。そして、この単語「首相」に対する分野<経済>の関連度（20.8）はその閾値より大きいので、単語「首相」は複数分野語と判定される。したがって、複数分野語リスト55には、図15に示すように、単語「首相：名詞」と共に、これに対応する強関連分野として<政治>及び<経済>が登録される。このような処理を関連度テーブルにおける全単語について行うことにより、図15に示すような複数分野語リストが完成する。図15のリストは、例えば単語「政治：名詞」は分野<政治>及び<経済>に関連が強く、単語「市場：名詞」に分野<経済>及び<国際>に関連が強いことを示している。

【0071】このようにして作成された複数分野語リスト55は、分類用情報記憶部9に記憶され、未分類文書の分類のために利用される。

【0072】なお、この実施の形態では、前述のように強関連分類の判定のための閾値を固定値とはせず、各単語の関連度の最大値に合わせて求めることとしたので、各単語の関連度の分布に合わせ、相対的にみて関連度が高い分野を強関連分野として抽出することができる。すなわち、例えば学習に用いた分類済み文書群においてたまたま出現回数が少なかったような単語についても、その単語の強関連分野を適切に求めることができる。

【0073】このようにして複数分野語リスト55が完成すると、次に複数分野語処理部8は、S4にて、それら複数分野語を考慮して、未分類の文書を分類する際の分類基準となる分類用情報を作成する。ここでは、分類用情報として、前述の関連度テーブル54を複数分野語を考慮して詳細化した詳細化関連度テーブル56と、各複数分野語についての共起単語の傾向を示す共起ベクトル57を作成する。以下、詳細化関連度テーブル56の作成手順及び共起ベクトル57の作成手順を順に説明する。なお、詳細化関連度テーブル56の作成と共起ベクトル57の作成はいずれを先に行ってもよい。

【0074】まず、図6を用いて詳細化関連度テーブル56の作成手順を説明する。詳細化関連度テーブル56は、例えば、図19に示すように、複数分野語を（単語、強関連分野）の組合せごとに別々の単語と捉えて分割（例えば「首相」を「首相（政治）」と「首相（経済）」に分割）し、関連度を求め直した結果である。このテーブルの作成手順は、大まかに分けて、複数分野語を（単語、強関連分野）の組合せごとに別々の単語とみなして頻度集計テーブル53（例えば図13参照）を詳細化するプロセスと、この結果得られた詳細化頻度集計テーブルから詳細化関連テーブルを作成するプロセスと、を含んでいる。図6では、S401～S405が前者のプロセスに対応し、S406～S407が後者のプ

ロセスに対応する。

【0075】すなわち、まず複数分野語処理部8は、S401にて、頻度集計テーブル53に登録された全単語について処理が完了したか検査する。完了していない場合には、S402にて頻度集計テーブル53から未処理の単語を1つ取り出す。そして、S403にて、選択した単語が複数分野語リスト55に登録された単語であるか否か（すなわち複数分野語か否か）チェックする。その単語が、複数分野語リスト55に登録されていた場合は、複数分野語処理部8は、S404にて、当該単語（すなわち複数分野語）と、これに対応する複数の強関連分類との各組合せごとに、詳細化頻度集計テーブルに欄を作成し、頻度集計テーブル53における当該単語の各分野での頻度値を、それら各組合せの欄に配分する。

【0076】すなわち、図18に示すように、複数分野語（例えば「首相」）を各強関連分野（「首相」については＜政治＞、＜経済＞）ごとに別々の単語（単語「首相（政治）」及び単語「首相（経済）」）と捉え、これら新たな単語について詳細化頻度集計テーブルに欄を作成する。そして、このようにして作成した欄「単語（分野）」における各分野に対し、頻度集計テーブルの頻度値を配分していく。配分する頻度値は次のように決定する。まず、当該「単語（分野）」の（）内の「分野」に一致する分野については、頻度集計テーブル53におけるその分野の頻度値をそのまま設定する。また、当該「単語（分野）」の（）内の「分野」には一致しないが、当該「単語」の強関連分野の一つである分野については、その頻度を0とする。そして、当該「単語」の強関連分野以外である分野については、頻度集計テーブル53におけるその分野についての頻度を当該単語の強関連分野の数で割り、その結果を設定する。この配分の仕方を具体的に示すと次のようになる。

【0077】図13の頻度集計テーブルから図18の詳細化頻度集計テーブルを作成する場合を例にとって説明する。複数分野語処理部8は、図13のテーブルから「首相」の各分野についての頻度値を取り出し、これらを順番に図18のテーブルの「首相（政治）」及び「首相（経済）」に配分していく。まず、図13における「首相」の＜政治＞に対する頻度50は、＜政治＞は「首相」の強関連分野なので、（）内の分野の一致する「首相（政治）」にそのすべてが配分され、（）内の分野の異なる「首相（経済）」には全く配分されない。したがって、図18の詳細化頻度集計テーブルの分野＜政治＞においては、「首相（政治）」の頻度は50、「首相（経済）」の頻度は0と設定される。逆に、同じく「首相」の強関連分野である＜経済＞については、図13における頻度30がすべて「首相（経済）」に対して配分され、「首相（政治）」の頻度は0に設定される。また、図13における「首相」の＜労働＞に対する頻度3は、＜労働＞は「首相」の強関連分野ではないので、

これを「首相」の強関連分野の数2で割った結果の1.5が「首相（政治）」と「首相（経済）」とに等しく配分される。したがって、図18のテーブルの分野＜労働＞においては、「首相（政治）」、「首相（経済）」が共に頻度1.5となる。

【0078】一方、S403で、選択した単語が複数分野語リスト55にない単語と判定された場合は、複数分野語処理部8は、S405にて、頻度集計テーブル53における当該単語の各分野についての頻度を、そのまま詳細化頻度集計テーブルに設定する。

【0079】例えば、図13の頻度集計テーブルにおける単語「薄商い」は、複数分野語ではないので、頻度集計テーブルにおける「薄商い」の各分野についての頻度が、そのまま詳細化頻度集計テーブルにおける「薄商い」の各分野についての頻度として設定される。

【0080】このような処理を頻度集計テーブル53の全登録単語に対して行うことにより、図18に示すように、複数分野語を強関連分類語ごとに分割することにより詳細化された詳細化頻度集計テーブルが完成する。

【0081】詳細化頻度集計テーブルが完成すると、次に複数分野語処理部8は、このテーブルに対して前述の式（1）を適用して理論頻度を算出する（S406）。そして、S2の場合と同様に、この理論頻度の算出結果を用いてカイ2乗検定の応用である前述の式（2）の演算を行うことにより、詳細化頻度集計テーブルにおける単語、分野の各組合せごとに、それら両者の関連度を算出する。この結果、複数分野語を強関連分類語ごとに別々の単語に分割した場合における、各単語と各分野との関連度を示した詳細化関連度テーブルが完成する。

【0082】図19は、このようにして作成された詳細化関連度テーブルの一例である。図19のテーブルは、図18の詳細化頻度集計テーブルから作成されたものである。図19の詳細化関連度テーブルを図14の関連度テーブルと比較すると、例えば図14における「首相」の＜政治＞に対する関連度と比較した場合、図19の「首相（政治）」の＜政治＞に対する関連度はそれより大きくなっており、「首相（経済）」の＜政治＞に対する関連度はそれより小さくなっているのが分かる。逆に、図14における「首相」の＜経済＞に対する関連度と比較した場合、図19の「首相（政治）」の＜経済＞に対する関連度はそれより小さくなり、「首相（経済）」の＜経済＞に対する関連度はそれより大きくなっている。

【0083】このようにして作成された詳細化関連度テーブル56は、分類用情報記憶部9に格納される。

【0084】次に共起ベクトル57の作成手順について説明する。この実施の形態では、すでに説明したように、複数分野語についてのみ共起ベクトルを作成する。しかも、この共起ベクトルは、当該複数分野語の強関連分野を考慮して、複数分野語と強関連分野との組合せご

とに作成する。以下、図7を参照して、共起ベクトルの作成手順を説明する。

【0085】まず、複数分野語処理部8は、S411にて、全分類済み文書に対して処理が完了したかを検査する。完了していない場合には、S412にて、分類済み文書の一つを選択し、その文書の単語分割情報と頻度情報を分類済み文書単語分割／頻度情報記憶部4から取り出す。次に、複数分野語処理部8は、取り出した頻度情報にあるすべての単語について処理が終わったかをS413で検査する。終わっていない場合は、S414にて、頻度情報から単語の一つを選択する。そして、S415では、複数分野語処理部8は、選択した単語が複数分野語かどうかを複数分野語リスト55を参照して判定する。ここで、複数分野語でない場合は、その単語については何もせず、S413に戻る。

【0086】S415において当該単語が複数分野語であった場合は、複数分野語処理部8は、S416において、現在処理中の文書の分類先の分野を文書分類先テーブル52にて調べ、その分野が複数分野語リスト55における当該単語の強関連分野に含まれているかどうかを判定する。

【0087】この判定の結果、現在処理中の文書の分類先が、当該単語の強関連分野に含まれていなければ、当該単語については何も処理を行わずにS413に戻る。一方、S416において、現在処理中の文書の分類先が、当該単語の強関連分野に含まれている場合には、複数分野語処理部8は、S417にて、当該文書の単語分割情報を参照して、例えば同一段落などの所定の範囲内において当該単語と共起する単語を当該文書全体にわたって調べる。この時、これら各共起単語の頻度も同時にカウントする。そして、複数分野語処理部8は、検出した共起単語とその頻度とを、S418にて共起頻度情報に反映させる。このようにして1つの単語についての処理が終わるとS413に戻って次の単語の処理に移る。

【0088】このような処理よれば、各複数分野語ごとに、当該複数分野語の強関連分野に属する文書において当該複数分野語と共起した単語及びその頻度の傾向が、共起頻度情報として求められる。

【0089】図16は、図12の単語分割情報及び図15の複数分野語リストに従って作成された共起頻度情報の内容を模式的に示した図である。図16では、「首相」、「市場」などの各複数分野語に対して、それぞれ対応する強関連分野が関連付けられており、さらにその複数分野語と強関連分野との組合せに対して、それぞれ共起単語及びその頻度を含んだデータのリストが関連付けられている。図16は、例えば、「総裁」という単語が、＜政治＞分野の文書では複数分類語「首相」に対して5回共起し、＜経済＞分野の文書では複数分類語「首相」に対して3回共起したことを示している。

【0090】ここまでの処理の具体例を以下に示す。ま

ず、図9の文書がS412で選択されたとする。この場合、複数分野語処理部8は、この文書に対応する頻度情報及び単語分割情報として、図11及び図12に示した情報を分類済み文書単語分割／頻度情報記憶部4から取り出す。そして、頻度情報の全ての単語について処理が終わるまでS413以下の処理を繰り返す。例えば、S414で図11から「会見：サ変名詞」が選択されたときには、この単語は図15の複数分野語リストに存在しないため、S415の判定により何も行わずにS413に戻る。順次処理が進んで、S413で「首相：名詞」が選択されたときには、S415でこの単語は図15の複数分野語リストに存在することを検知し、S416に進む。S416では、図10の文書分類先テーブルから、図9の文書の分類先が＜政治＞分野であることを検知するとともに、複数分野語リストから当該複数分野語「首相：名詞」の強関連分野が＜政治＞及び＜経済＞であることを検知し、これらのことから、当該文書の分類先の分野が当該複数分野語の強関連分野の一つであることを検知する。したがって、S416の判定結果はYESとなり、以下、当該文書の分類先分野と当該複数分野語の強関連分野との一致点である＜政治＞分野について、S417及びS418の処理が行われる。この具体例では、S417にて、現在処理中の複数分野語と同じ段落に出現したものを共起単語として抽出する。したがって、図9の段落70からは、複数分野語「首相：名詞」の共起単語として、「内閣：名詞」、「支持率：名詞」、「理由：名詞」、「海部：名詞」、「続投論：名詞」、「三塚：名詞」、「反発：サ変名詞」が抽出される。また、これら各共起単語の同段落70における頻度（いずれも1である）も検出される。そして、S418では、図16に示す共起頻度情報において、「首相：名詞」の＜政治＞に関連付けられた各共起単語の頻度にそれぞれS417で検出された頻度を加える。このような処理をすべての単語について繰り返すことにより、図16に示すような共起頻度情報が得られる。

【0091】なお、この例では、同一段落に出現したものを共起単語としたが、これに限らず、広く同一文書に出現したもののすべてを共起単語としてもよいし、逆に範囲を狭め、同一の文に出現したもののみを共起単語としてもよい。この実施の形態では、単語分割情報（たとえば図12）を作成しているの、このようないずれの場合にも対応することができる。

【0092】このようにして共起頻度情報が完成すると、複数分野語処理部8は、この共起頻度情報から、各複数分野語・強関連分野の組合せごとについて、共起ベクトルを作成する。共起ベクトルは、共起頻度情報（例えば図16）における共起単語の頻度値をその共起単語に対応する各基底の成分値とするベクトルである。ただし、この実施の形態では、学習に用いた分類済み文書群に現れたすべての単語を共起ベクトルの基底とし、共起

しなかった単語に対する成分値は0とすることで共起ベクトルの基底を統一している。例えば、図16の共起頻度情報から求めた「首相（政治）」、「首相（経済）」の共起ベクトルは、それぞれ図17の（a）に示すようなものとなる。

【0093】そして、複数分野語処理部8は、S419で、各共起ベクトルを長さ1に正規化することにより、共起ベクトルの長さの差を吸収する。すなわち、以下では、共起ベクトルの「方向」のみについて注目する。この方向が、複数分野語と強関連分野との組合せである「単語（分野）」に対する共起単語の出現傾向を表す。例えば、図17の（a）に示した各共起ベクトルは、同図の（b）のように正規化される。

【0094】このようにして作成された各複数分野語と強関連分野との各組合せについての共起ベクトルは、分類用情報記憶部9に格納される。

【0095】これで、分類用情報記憶部9には、分類対象文書の分類処理の際の基準となる複数分野語リスト55、詳細化関連度テーブル56及び各共起ベクトル57がすべて用意された。以下、これらの情報を用いた文書の分類処理の各ステップ（図2のS5～S7）の処理手順を更に詳細に説明する。

【0096】まず、分類対象文書が与えられた場合、まずS5にて、単語分割／頻度抽出処理部3が、図3に示した処理手順により当該分類対象文書の出現単語を解析し、当該分類対象文書の単語分割情報及び頻度情報を作成する。この手順は、分類済み文書の場合と同様なので説明は省略する。

【0097】例えば、分類対象文書として図20に示す文書が与えられたとする。この文書は、本来＜経済＞に分類されるべき文書である。単語分割／頻度抽出処理部3は、この文書を解析し、図21に示す頻度情報及び図22に示す単語分割情報を作成する。例えば、図21は、図20の文書に、「東京証券取引所：名詞」が1回、「首相：名詞」が3回出現していることを示している。また、図22は、例えば、「首相：名詞」が文書冒頭から223バイト目に出現したことを示している。

【0098】このようにして得られた分類対象文書の分類対象文書の単語分割情報及び頻度情報（単語分割／頻度情報58）は、分類対象文書単語分割／頻度情報記憶部5に格納される。

【0099】分類対象文書の単語分割情報及び頻度情報が得られると、次に複数分野語処理部8及び分類先決定部10により、当該分類対象文書の特徴を表す文書ベクトル60を作成し（S6）、この文書ベクトル60に基づき分類先を決定する（S7）。この一連の処理の詳細な手順を、図8を参照して説明する。なお、図8における各ステップは、S61～S66が図2のS6に対応し、S70が図2のS7に対応する。

【0100】分類先決定部10は、分類対象文書の頻度

情報（例えば図21）を取り出し、この頻度情報にある単語を先頭から順に一つずつ取り出して処理していく。このため、まずS61にて、その頻度情報にあるすべての単語に対して処理が終わったかを検査する。終わっていない場合は、S62で、頻度情報における次の未処理単語を選択する。そして、S63では、分類用情報記憶部9に格納された共起ベクトルの情報から、選択した単語が共起ベクトルを持つ単語かどうかを判定する。なお、この判定は、当該単語が複数分野語リスト55に含まれるかどうかに基づき行ってもよい。

【0101】S63の判定の結果、当該単語が共起ベクトルを持たない単語であれば、分類先決定部10は、頻度情報における当該単語の頻度を、そのまま文書ベクトルにおける当該単語に対応する成分に設定する。

【0102】ここで、文書ベクトルは、学習に用いた分類済み文書群に現れる全単語を基底とするベクトルであり、基本的には、分類対象文書に出現した単語の頻度をその単語に対応する基底の成分値とするベクトルである（したがって、出現しなかった単語についての成分は0となる）。このため、S66では、分類対象文書における単語の頻度値を文書ベクトルに設定する。例えば、図21の頻度情報における「東京証券取引所：名詞」という単語は、共起ベクトルを持たないので、その単語の頻度1が、そのまま文書ベクトルにおける「東京証券取引所：名詞」の成分に設定される。

【0103】ただし、文書ベクトルでは、複数分野語については各強関連分野との組合せごとに1単語（例えば、分野付きの単語「首相（政治）」など）とみなし、各組合せをそれぞれ基底としている。すなわち、この文書ベクトルは、基本的には分類対象文書における単語の出現傾向を示すベクトルであるが、更に複数分野語がどの分野（強関連分野）の単語として出現したかを示す情報を含んだものとなっている。

【0104】このような文書ベクトルを作成のため、分類対象文書内に出現した複数分野語については、その頻度を各強関連分野との組合せごとに分配する必要がある。このための処理がS64及びS65の各ステップである。

【0105】すなわち、S63の判定で共起ベクトルを持つ単語（すなわち複数分野語）と判定された場合、S64では、まず複数分野語処理部8にて、当該単語の文書共起ベクトルを作成する。ここで、文書共起ベクトルは、前述の共起ベクトルと同様、文書の所定範囲内（例えば同一段落内）で当該単語と共起した単語の頻度を各成分値としたベクトルである。共起ベクトルと文書共起ベクトルとの相違は、前者は学習に用いた複数の分類済み文書から作成されたものであるのに対し、後者は1つの分類対象文書のみから作成されたものである点である。すなわち、文書共起ベクトルは、複数分野語が、分類対象文書においてどのような単語と共起したかという

共起単語の傾向を表すベクトルである。

【0106】S63で共起ベクトルを持つと判定された単語についての文書共起ベクトルの作成は、共起ベクトルの作成手順（図7参照）におけるS417及びS418とほぼ同様の処理にて行うことができる。すなわち、単語分割情報（例えば図22）を参照して当該単語の共起単語を検出し、それら共起単語の頻度をそれぞれ対応する成分の値として設定すればよい。

【0107】例えば、図20の分類対象文書に出現する単語「首相：名詞」は、図17などに示すように共起ベクトルを持つ複数分野語である。この図20の文書から作成した「首相：名詞」の文書共起ベクトルを図23に示す。図23の文書共起ベクトルは、共起の範囲を同一段落内に限った場合の例である。図23と図17を比較すれば分かるように、文書共起ベクトルは、共起ベクトルと同じ基底のベクトルである。なお、この文書共起ベクトルの長さを正規化する必要はない。

【0108】選択した単語についての文書共起ベクトルが求められると、S64では、更に分類先決定部10が、分類用情報記憶部9から当該単語についての各共起ベクトルを取り出し、文書共起ベクトルとこれら各共起ベクトルとの内積をそれぞれ計算する。得られた各内積値は、当該単語（この単語は複数の分野に強い関連を持つ複数分野語である）が、当該分類対象文書においてはどの分野の単語として出現している可能性が高いかを示している。

【0109】すなわち、各共起ベクトルは、複数分野語と強関連分野との組合せについて求められており、例えば「首相（政治）」についての共起ベクトルは、「首相」という複数分野語が＜政治＞分野の文書に出現する場合にはどのような単語と共起しているかという傾向を表している。したがって、分類対象文書における当該単語の共起単語の傾向を表す文書共起ベクトルと、各分野の文書における当該単語の共起単語の傾向を表す各共起ベクトルとを比較すれば、それらの類似度合いから分類対象文書において当該単語はどの分野の単語として現れているかを知ることができる。

【0110】共起ベクトルはすべて長さ1に正規化されているので、文書共起ベクトルと共起ベクトルの内積値の大小は、純粋に両ベクトルの方向の類似性のみを示し、内積値が大きいほど両ベクトルの方向は類似しているといえる。文書共起ベクトルや共起ベクトルの方向は、共起単語の出現の傾向を表しているので、これら両ベクトルの内積値を、共起単語の出現傾向の類似度合いを示していると捉えることができる。したがって、この実施の形態では、文書共起ベクトルと各共起ベクトルの内積値を、当該単語がどの分野の単語であるかを示す指標値として用いる。例えば、図23の「首相」の文書共起ベクトルと、図17（b）の「首相（政治）」の共起

「首相」が分野＜政治＞の単語である可能性の大きさを示している。同様に、図17（b）の「首相（経済）」の共起ベクトルとの内積値は、「首相」が＜経済＞の単語である可能性の大きさを示している。

【0111】そこで、分類先決定部10は、S65にて、求められた各内積値の大きさの比率に従って、頻度情報（図21参照）における当該単語の頻度値を、当該単語と強関連分野との各組合せに対して比例配分し、その配分結果を文書ベクトルに設定する。

【0112】例えば、前述の単語「首相」についての例を用いて説明すると、「首相」の文書共起ベクトル（図23）と、「首相（政治）」「首相（経済）」についての各共起ベクトル（図17）との内積値は、それぞれ0.82及び4.08となるので、分類対象文書における「首相」の頻度3を、それら内積値の比率に応じて比例配分すると、「首相（政治）」に配分される頻度値は、

【数4】

$$\frac{0.82}{(0.82 + 4.08)} \times 3 = 0.5$$

となり、「首相（経済）」に配分される頻度値は、

【数5】

$$\frac{4.08}{(0.82 + 4.08)} \times 3 = 2.5$$

となる。したがって、S65では、文書ベクトルの「首相（政治）」及び「首相（経済）」の成分値に、0.5及び2.5をそれぞれ設定する。

【0113】以上説明したS61～S66の処理を、分類対象文書の頻度情報のすべての単語について繰り返すことにより、当該分類対象文書の文書ベクトルが求められる。例えば、図24は、分類対象文書の頻度情報（図21）の各単語の頻度を、各単語の共起ベクトル及び文書共起ベクトルを使って分割した結果を示す例である。この表における各単語の頻度を、所定の順番（図19等）に示す詳細化関連度テーブルでの単語の登録順序）に合わせて並べたものが、文書ベクトルとなる。このようにして作成された文書ベクトルは、分類対象文書の出現単語の傾向を表している。

【0114】分類対象文書の文書ベクトルが完成すると、分類先決定部10は、S70にて、この文書ベクトルと分類用情報記憶部9の詳細化関連度テーブルとを用いて、当該分類対象文書の分類先を決定する。このため、分類先決定部10は、次の式を用いて当該分類対象文書の各分野*i*への関連度*S_i*を計算する。

【0115】

【数6】

$$S_i = \sum_{j=1}^L (Y_{ij} \cdot D_j) \quad (i=1, \dots, N)$$

この式で、*D_j*は文書ベクトルにおける単語*j*に対応す

る成分値を示し、 L_i は複数分野語を各強関連分野ごとに別々の単語（例えば「首相（政治）」、「首相（経済）」）とみなしたとき単語の総数を示す。この演算は、詳細化関連度テーブル（例えば図19）を各分野ごとに分割し、各分野の列をそれぞれ当該分野の単語の出現傾向を示すベクトルとみなし、このベクトルと文書ベクトルとの内積を求める演算と捉えることができる。

【0116】例えば、図20の文書の各分野に対する関連度は、図19の詳細化関連度テーブルと図24の表に対応する文書ベクトルとに基づき、図25に示すような内積演算で求められる。図25に示すように、図20の文書は、分野<経済>に対する関連度が最も大きくなっている。

【0117】分類先決定部10は、例えば関連度 S_i が最大となる分野を、分類対象文書の分類先に決定する。なお、分類先を一つに限らずに、例えば関連度 S_i が所定順位内の分野を分類先としてもよい。また、関連度 S_i が最大となる分野を分類先として抽出するのに加え、その関連度の最大値に対して所定の割合以上の関連度を持つ分野を副分類先として抽出するような応用も考えられる。得られた文書分類結果61は、文書分類結果記憶部11に格納される。

【0118】図25の例では、図20の文書が分野<経済>に正しく分類されていることが分かる。

【0119】これに対し、図26は、図20の文書の分類先を、複数分野語を分割していない図14の関連度テーブル及び図21の頻度情報から求めた場合の演算例を示している。図26の例では、分野<政治>に対する関連度が最大値となっており、これでは図20の文書が、本来分類されるべき<経済>ではなく、<政治>に誤って分類されてしまう。このように、複数分野語に注目しない従来の手法では、分類先を誤る可能性が高いことが分かる。これに対し、この実施の形態の手法は、複数分野語に注目して、複数分野語を各強関連分野ごとに別々の単語に分割して取り扱うことにより、文書における単語の出現傾向をより詳細に分析することができるので、より好ましい分類結果を得ることができる。

【0120】以上説明したように、この実施の形態1によれば、分類済み文書を解析することにより複数分野語を抽出し、この複数分野語に注目して、分類の際の基準となる関連度テーブルや、分類対象文書の出現単語の頻度情報を詳細化することにより、分類対象文書の各分野への関連度をより詳細に分析することができるので、類似する分野（共通の単語がある程度以上の頻度で出現する複数の分野）間での分類の精度を向上させることができる。したがって、この実施の形態1は、分類を細かくして類似する分野が多くなるような場合においても、分類精度の劣化を抑えることができる。

【0121】また、この実施の形態1は、これら関連度テーブルなどの詳細化を、複数分野語についての共起単

語の傾向を示すベクトル（共起ベクトル、文書共起ベクトル）に基づき行うことにより、複数分野語が現れた周囲の状況に基づき当該複数分野語がどの分野の単語として現れたかを評価することができる。これにより、従来技術において誤分類の原因となっていた複数分野語を、各強関連分野ごとに別々の単語として切り分けて取り扱うことが可能となるとともに、複数分野語の出現状況を踏まえて適切にその切り分けを行うことができる。

【0122】実施の形態2. 次に、この発明の実施の形態2を説明する。この実施の形態2は、各単語の上位概念語が登録されたシソーラスを利用して、共起ベクトルや文書共起ベクトルを上位概念語を含めた形式に拡張することにより、複数分野語の頻度の分割をより適切に行おうとするものである。

【0123】この実施の形態2のシステムの構成を図27に示す。図27において、図1と同様の構成要素には同一の符号を付してその説明を省略する。図27から分かるように、この実施の形態2の各構成要素は、シソーラス利用型複数分野語処理部21とシソーラス22以外は、図1の構成要素と同一である。シソーラス22には、各単語の意味概念の階層関係の情報を格納している。例えばシソーラス22は、単語「テニス」の上位概念として単語「球技」があり、単語「球技」の上位概念として単語「スポーツ」があるなど、各単語の階層的な上下関係の情報を記憶している。シソーラス利用型複数分野語処理部21は、図1（実施の形態1）における複数分野語処理部8に対応するものであり、基本的な処理動作は共通している。ただし、シソーラス利用型複数分野語処理部21は、共起ベクトルや文書共起ベクトルを作成する際に、このシソーラス22を利用して各共起単語の上位概念語を抽出し、この上位概念語をそれらベクトルに反映させる点が、図1の複数分野語処理部8と異なる。

【0124】この実施の形態2のシステムは、シソーラス利用型複数分野語処理部21における共起ベクトル及び文書共起ベクトルの作成手順のみが実施の形態1のシステムと異なるだけで、その他の処理手順は、実施の形態1の場合と全く同じでよい。したがって、以下では、実施の形態1と異なる部分である、シソーラス利用型複数分野語処理部21における共起ベクトルの作成手順のみを詳細に説明する。

【0125】図28は、このシソーラスを利用した分類学習時の共起ベクトルの作成手順を示すフローチャートである。図8（実施の形態1での共起ベクトル作成処理）におけるステップと同じ処理を示すステップについては、図25においても図8と同じ符号を付してその説明を省略する。

【0126】図28において、分類済み文書の頻度情報から複数分野語を取り出し、この複数分野語の共起単語を単語分割情報から抽出するまで（S417まで）の処

10

20

30

40

50

理は、実施の形態1と同様の処理である。このようにして、複数分野語の共起単語が取り出されると、シソーラス利用型複数分野語処理部21は、各共起単語の上位概念語をシソーラス22から取り出す。取り出した上位概念語には、共起単語の頻度に対し階層差に応じた重みを乗じた値を頻度値として割り当てる。そして、S502では、共起頻度情報（例えば図16参照）に対し、各共起単語の頻度値を反映させる。このとき、共起単語だけでなく、それらの上位概念語も共起頻度情報のリストに加え、その頻度値を反映させる。

【0127】例えば、S417にて、共起単語として「テニス」が抽出され、シソーラス22には図29に示すような単語の階層関係が記憶されていたとする。この場合、S501では、「テニス」の上位概念語として「球技」及び「スポーツ」が抽出される。ここでは、上位概念語には、例えば、共起単語から階層が1つ上に上がるごとに所定の重みを乗じた頻度値を割り当てる。例えば、共起単語「テニス」の頻度が1であり、前記所定の重みを0.5とすると、S501では、「テニス」「球技」「スポーツ」の各単語に対し、図29に示すようにそれぞれ1, 0.5, 0.25の頻度値が割り当てられる。そして、これら頻度値が、S502で共起頻度情報における対応単語の頻度値に加算される。そして、この共起頻度情報から共起ベクトルが作成され、長さ1に正規化される。

【0128】また、この実施の形態2では、文書共起ベクトルにも、上記共起ベクトルの場合と同様に、共起単語の上位概念語を反映させる。

【0129】すなわち、前記実施の形態1では、図30の(a)に示すように共起ベクトルには共起単語「テニス」そのものの頻度しか反映させていなかったが、この実施の形態2では、共起ベクトルに対し共起単語の上位概念語「球技」及び「スポーツ」の頻度も反映させる（なお、図30では、共起ベクトルの正規化は行っていない）。

【0130】分類済み文書の数は有限なので、分類学習時において、ある複数分野語の共起単語として現れる単語もある程度限られてくる。したがって、これから分類しようとする分類対象文書に、そのような限られた単語そのものが当該複数分野語の共起単語として現れるとは限らない。例えば、分類対象文書に、当該複数分野語の共起単語として、分類学習時の共起単語そのものは現れなかったが、それに類似した単語は現れたというような場合も考えられる（実際に、類似した単語が現れる確率は高いと考えられる）。このような場合、分類学習時に実際に共起した単語のみしか共起ベクトルに考慮しないとすれば、分類対象文書における複数分野語の頻度のある強関連分野に配分する際に、たまたまその分類学習時における当該強関連分野の共起単語そのものが現れなかっただけのために、その強関連分野に対する配分比率が

下がってしまうということもあり得る（類似の単語は出現しているのだから、実際にはその複数分野語はその強関連分野に対してもっと関連が深いはずである）。このような誤りが生じるのは、分類対象文書にせっかく現れた類似単語の情報が捨てられてしまうためといえる。

【0131】これに対し、この実施の形態2では、共起単語そのものだけでなく、その類似単語の一種である上位概念語をも共起ベクトル、文書共起ベクトルに反映させるので、分類対象文書内にある複数分野語に対する分類学習時の共起単語そのものが現れない場合でも、その上位概念語が共起単語として現れていれば、当該複数分野語が関連の深い分野を大きく誤るようなことはない（すなわち、複数分野語の頻度の配分比率をある程度修正することができる）このようにして求められる共起ベクトルや文書共起ベクトルは、分類対象文書に現れた複数分野語の頻度値を、当該複数分野語の強関連分野に配分する際（図8のS64、S65）に用いられる。そして、この結果求められた文書ベクトルを用いて分類対象文書の分類先が決定される。この実施の形態2では、共起ベクトル及び文書共起ベクトルは、実施の形態1と比べて詳細化されているが、このほかの詳細化関連度テーブルや文書ベクトルは実施の形態1のものと同じでよい。

【0132】このように、実施の形態2によれば、共起ベクトルのスパースネス（基底の疎さ）を吸収することが可能となり、実施の形態1で共起ベクトルでうまく処理できなかった分類対象文書に対しても、より好ましい分類結果を得ることができる。

【0133】以上、この発明の好適な実施の形態を説明した。以上に説明した各実施の形態は、あくまで一例に過ぎず、こほかにも様々なバリエーションが考えられる。

【0134】例えば、上記各実施の形態においては、文書から単語を切り出すために形態素解析を用いたが、この形態素解析の代わりに、平仮名、カタカナ、漢字、数字などの文字タイプの情報を利用して疑似的に単語分割を行うことも可能である。また、このほかにも言語処理分野で知られている様々な単語分割手法を適用することができる。

【0135】また、上記各実施の形態では単語と分野との関連度を、カイ2乗検定を応用した計算式により求めたが、関連度の求め方はこれに限らない。例えばTF・IDF(term frequency times inverse document frequency)など、統計分野で知られている様々な手法を利用することができる。

【0136】また、上記各実施の形態では、共起ベクトルを作成する際に、共起単語の調査範囲を同一段落とするほか、文書全体に範囲を広げたり、逆にその単語を含む文に範囲を限定したりすることが可能であると説明し

たが、このほかにも、注目する単語の前後所定文字数以内の範囲といった限定の仕方も可能である。これは、単語分割情報における単語位置の情報を用いて実現することができる。

【0137】また、分類対象文書と各分野との間の関連度を求めるための方法は、上記実施の形態1に示したベクトルの内積を利用する方法に限らず、様々な距離計算アルゴリズムを利用することが可能である。

【0138】また、上記実施の形態では、共起ベクトルを長さ1に正規化した。次のような方法により、より共起ベクトルを特徴付けることもできる。すなわち、上記実施の形態1の方法により求められた各共起ベクトルにおいて、複数の共起ベクトルが、ある共通の単語に対して共に0でない成分値を有している場合は、その成分値をそれら共起ベクトルの数で除した値に置き換えるという方法である。例えば、図31に示すように、実施の形態1の方法で(a)に示すような共起ベクトルが得られたとする。(a)の2つの共起ベクトルは、「総裁：名詞」に対応する成分値が共に0でない値となっているので、これを(b)に示すようにそれぞれ2で割った値に置き換えるのである。ある単語が複数の共起ベクトルにおいて0でない成分値を持つということは、その単語はそれら複数の共起ベクトルに対応する分野同士の間での分類においては重要度が低いと考えることができるので、このような方法により共起ベクトルに特徴づけを行うことができる。

【0139】

【発明の効果】以上説明したように、この発明によれば、分類済み文書から複数分野語を学習し、この複数分野語に注目して関連度テーブルや分類対象文書の単語の頻度情報を詳細化するので、分類対象文書の各分野への関連度をより詳細に分析することができ、類似する分野間での分類の精度を向上させることができる。したがって、この発明によれば、細かい分類を行う場合でも分類精度の劣化を抑えることができる。

【0140】また、単語の強関連分野の判定基準となる閾値を、関連度テーブルにおける当該単語の各分野に対する関連度の中の最大値に基づき定めるので、各単語の関連度の分布に合わせ、相対的にみて関連度が高い分野を強関連分野として抽出することができる。

【0141】また、分類対象文書の出現単語の頻度情報の詳細化を、複数分野語の共起単語の傾向を示す共起ベクトルを利用して行うことにより、複数分野語が現れた周囲の状況に基づき当該複数分野語がどの分野の単語として現れたかを評価し、適切に詳細化を行うことができる。

【0142】また、各単語の概念的な階層関係を記述したシソーラスを用いて共起ベクトルや文書共起ベクトルを拡張することにより、共起情報のスパースネスを解消してよりの確に単語が使われた状況を選択することがで

き、よりの確な文書の自動分類を行うことができる。

【図面の簡単な説明】

【図1】 この発明の実施の形態1のシステムの構成図である。

【図2】 この発明の実施の形態1のシステムにおける全体的な処理手順を示すフローチャートである。

【図3】 単語分割／頻度抽出処理部の処理手順を示すフローチャートである。

【図4】 関連度演算部の処理手順を示すフローチャートである。

【図5】 複数分野語処理部による複数分野語検出処理の手順を示すフローチャートである。

【図6】 複数分野語処理部による詳細化関連度テーブルの作成処理の手順を示すフローチャートである。

【図7】 複数分野語処理部による共起ベクトル作成処理の手順を示すフローチャートである。

【図8】 分類対象文書の分類の手順を示すフローチャートである。

【図9】 分類済み文書の一例を示す図である。

【図10】 文書分類先テーブルのデータ内容の一例を示す図である。

【図11】 分類済み文書の頻度情報の一例を示す図である。

【図12】 分類済み文書の単語分割情報の一例を示す図である。

【図13】 頻度集計テーブルの一例を示す図である。

【図14】 関連度テーブルの一例を示す図である。

【図15】 複数分野語リストの一例を示す図である。

【図16】 共起ベクトルの作成のために構築される共起頻度情報の一例を示す図である。

【図17】 共起ベクトルの一例を示す図である。

【図18】 詳細化頻度集計テーブルの一例を示す図である。

【図19】 詳細化関連度テーブルの一例を示す図である。

【図20】 分類対象文書の一例を示す図である。

【図21】 分類対象文書の頻度情報の一例を示す図である。

【図22】 分類対象文書の単語分割情報の一例を示す図である。

【図23】 文書共起ベクトルの一例を示す図である。

【図24】 分類対象文書の文書ベクトルに対応する各単語の頻度を示す表の一例を示す図である。

【図25】 図20の分類対象文書の各分野に対する関連度を、複数分野語を考慮した実施の形態1の手法により求めた結果を示す図である。

【図26】 図20の分類対象文書の各分野に対する関連度を、複数分野語を考慮しない手法で求めた結果を示す図である。

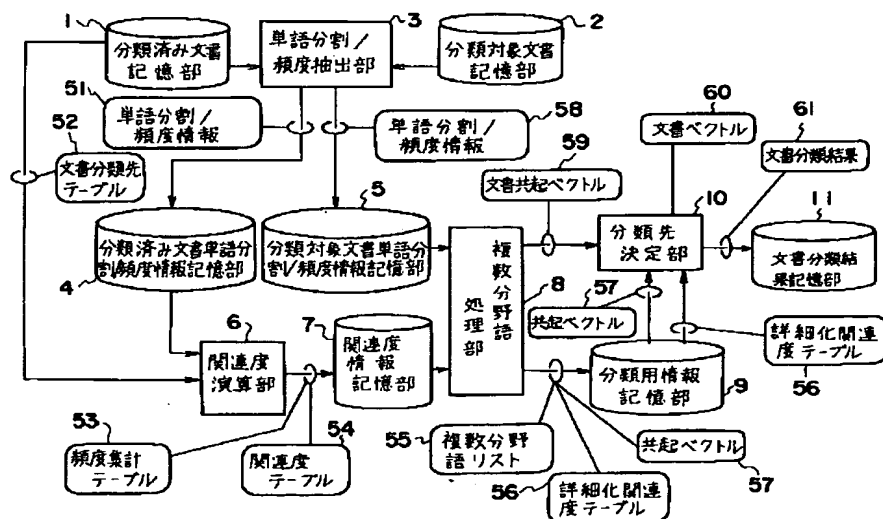
【図27】 この発明の実施の形態2のシステムの構成

【図36】 図34の従来装置で生成される文書ベクトルの一例を示す図である。

1 分類済み文書記憶部、2 分類対象文書記憶部、3 単語分割／頻度抽出部、4 分類済み文書単語分割／頻度情報記憶部、5 分類対象文書単語分割／頻度情報記憶部、6 関連度演算部、7 関連度情報記憶部、8 複数分野語処理部、9 分類用情報記憶部、10 分類先決定部、11 文書分類結果記憶部、21 シソーラス利用型複数分野語処理部、22 シソーラス、51, 58 単語分割／頻度情報、52 文書分類先テーブル、53 頻度集計テーブル、54 関連度テーブル、55 複数分野語リスト、56 詳細化関連度テーブル、57 共起ベクトル、59 文書共起ベクトル、60 文書ベクトル、61 文書分類結果。

20

【图 1-1】



単語	品詞	頻度

会見	サ変名詞	1
会長	名詞	2
海部	名詞	2
記者	名詞	1

国民	名詞	4

首相	名詞	3

総裁	名詞	3

内閣	名詞	3

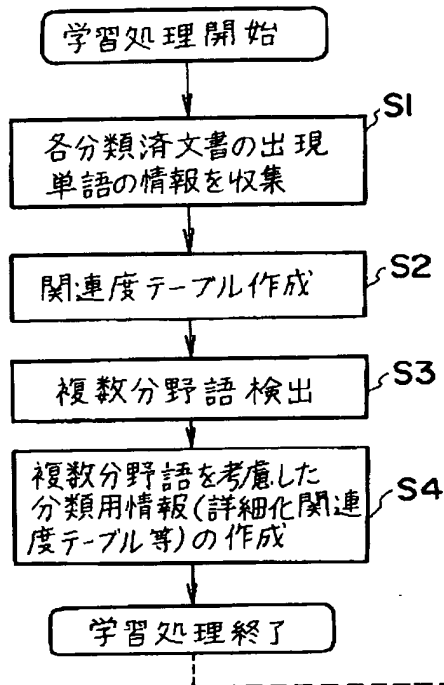
【图 10】

自民党三重県会長の三塚博氏は2日、長野市内で記者会見し、金丸信氏（竹下派会長）が調議で「国民の声を聞く必要がある」と賛同し、内閣支持率の高い郵政首相の続投支持を表明したことについて「昨年の総選挙で自民党が圧勝したこと示された国民の意思は（総選挙後も）政党の選択に任せておくことだ」と述べた。

文書名（ファイル名）	分野
...	
11/04M/04--09	政治
....
11/12M/09--06	経済
11/12M/09--08	政治、経済
....

【図2】

分類学習時の処理



【図12】

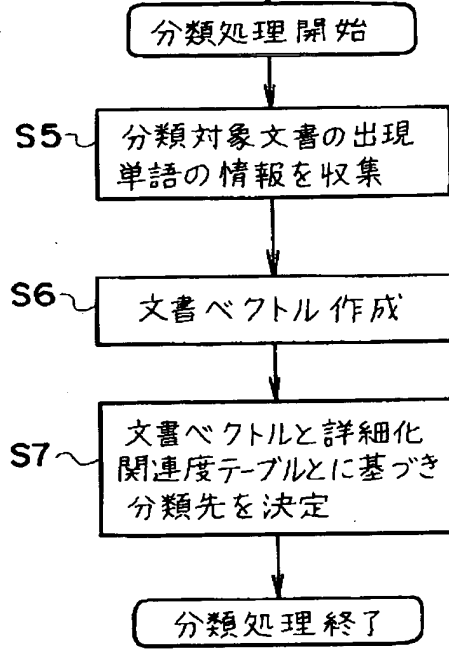
情報タイプ	単語位置	単語	品詞
P			
S			
W	3	内閣	名詞
W	7	支持率	名詞
W	13	理由	名詞
W	19	海部	名詞
W	23	首相	名詞
W	27	続投論	名詞
...			
P			
S			
W	55	自民党	名詞
W	61	三塚	名詞
W	67	会長	名詞
...			
W	99	記者	名詞
S			
W	544	首相	名詞
...			

【図23】

首相 = (2, 0, 0, 4, 1, ...)

内閣 国民 総裁 市場 経済
(名詞) (名詞) (名詞) (名詞) (名詞)

分類実行時の処理



【図13】

単語	政治	経済	労働	...	スポーツ	国際
首相	50	30	3		1	10
...						
出来高	1	10	0		0	2
得票い	0	5	0		0	0
市場	0	10	0		0	8
...						

【図14】

単語	政治	経済	労働	...	スポーツ	国際
首相	66.7	20.8	0.0		-1.6	1.6
...						
出来高	-3.2	9.0	-1.0		-1.0	0.0
得票い	-1.6	10.0	-0.3		-0.3	-0.6
市場	-5.0	9.0	-1.0		-1.0	18.0
...						

【図15】

単語	品詞	強関連分野
...		
首相	名詞	政治、経済
...		
市場	名詞	経済、国際
...		

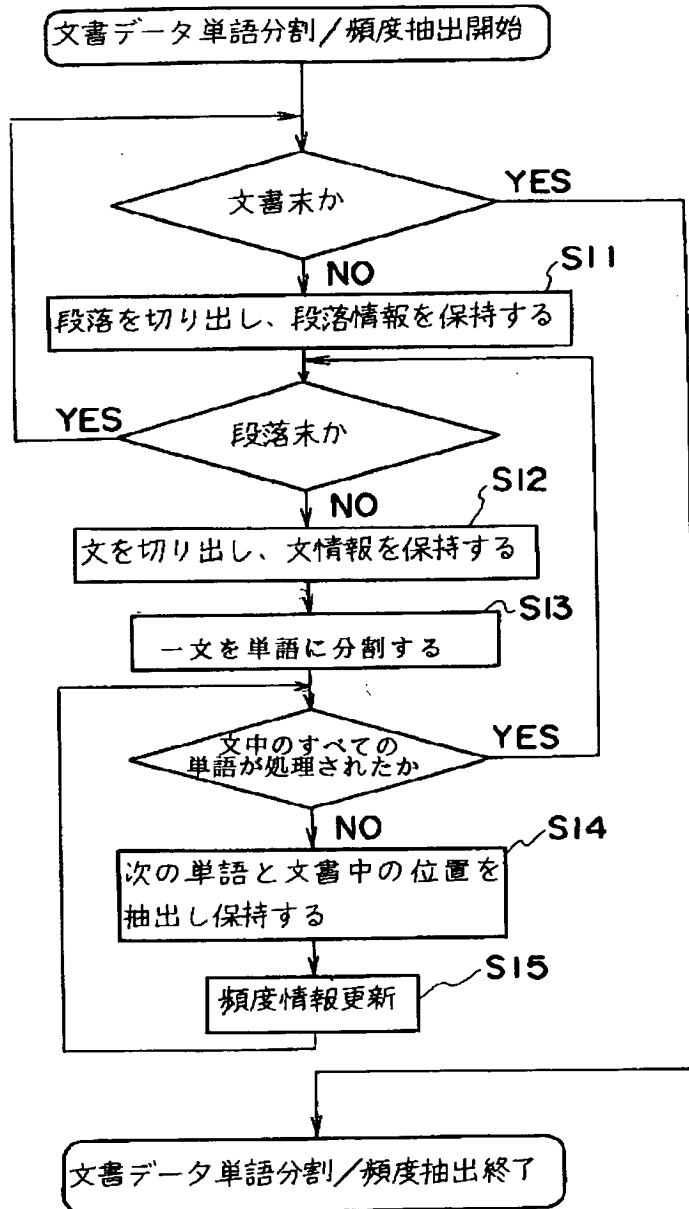
【図21】

単語	品詞	頻度
...		
東京証券取引所	名詞	1
下半期	名詞	1
各社	名詞	2
株価	名詞	1
市場	名詞	4
...		
首相	名詞	3
...		
出来高	名詞	8
...		
得票い	名詞	3
...		

【図24】

単語	品詞	頻度
得票い	名詞	3
市場(経済)	名詞	3.3
市場(国際)	名詞	0.8
...		
出来高	名詞	3
...		
首相(政治)	名詞	0.5
首相(経済)	名詞	2.5
...		

【図3】



【図19】

単語	政治	経済	労働	...	スポーツ	国際
首相(政治)	160.0	-8.0	-0.1		-1.1	0.2
首相(経済)	-6.6	114.0	0.0		-0.5	2.0
...						
出来高	-3.2	9.0	-1.0		-1.0	0.0
薄商い	-1.6	10.0	-0.3		-0.3	0.6
市場(経済)	-2.6	29.0	-0.5		-0.5	-1.0
市場(国際)	-2.3	-1.8	-0.4		-0.4	53.5
...						

【図18】

単語	政治	経済	労働	...	スポーツ	国際
首相(政治)	30	0	1.5		0.5	5
首相(経済)	0	30	1.5		0.5	5
...						
出来高	1	10	0		0	2
薄商い	0	5	0		0	0
市場(経済)	0	10	0		0	0
市場(国際)	0	0	0		0	8
...						

【図22】

情報タイプ	単語位置	単語	点
P			
S			
W	11	薄商い	名詞
W	27	出来高	名詞
W	51	東証	名詞
W	68	1部	名詞
P			
S			
W	68	東京証券取引所	名詞
W	77	市場	名詞
W	81	1部	名詞
W	85	出来高	名詞
...			
W	192	水準	名詞
S			
W	207	市場	名詞
...			
W	228	首相	名詞

【図25】

計算結果

政治: $160.0 \times 0.5 + (-6.6) \times 2.5 + (-0.1) \times 3 + (-1.1) \times 8 + (-2.6) \times 3 + (-2.3) \times 0.8 + \dots = 38.9 + \dots$

経済: $(-8.0) \times 0.5 + 114.0 \times 2.5 + 9.0 \times 3 + 10.0 \times 3 + 29.0 \times 3 + 5.0 \times (-1.0) \times 0.8 + \dots = 429.8 + \dots$

労働: $(-0.1) \times 0.5 + (0.0) \times 2.5 + (-1.0) \times 3 + (-0.3) \times 8 + (-0.5) \times 3 + (-0.4) \times 0.8 + \dots = -5.8 + \dots$

スポーツ: $(-1.1) \times 0.5 + (-0.5) \times 2.5 + (-1.0) \times 3 + (-0.3) \times 8 + (-0.5) \times 3 + (-0.4) \times 0.8 + \dots = -7.6 + \dots$

国際: $0.2 \times 0.5 + 2.0 \times 2.5 + 0.0 \times 3 + 0.6 \times 8 + (-1.0) \times 3 + 53.5 \times 0.8 + \dots = 42.9 + \dots$

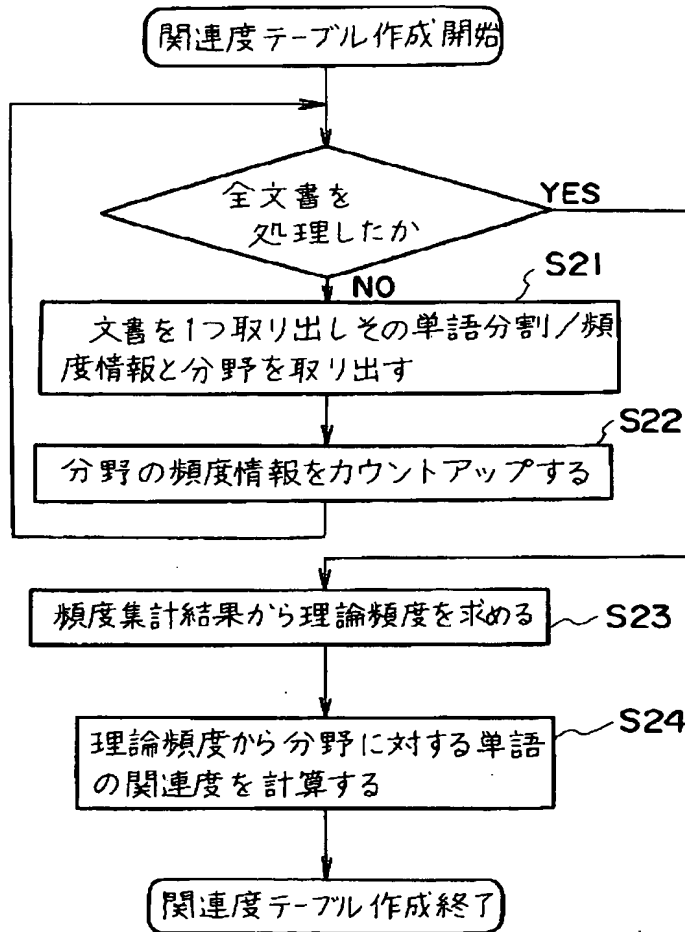
【図20】

記録的な薄商い 11日の出来高1億5000万株 東証1部

東京証券取引所市場1部の11日の出来高は、1億5437万株と1日の立ち合いでは1984年5月28日(1億3563万株)以来の低い水準となった。市場関係者は官託首相就任と同時に公定歩合引き下げがあると期待していたが、早くも今週後半になったため、市場の見送り気分が濃まった。8月中旬決算で赤字や大幅減益となった証券各社は、年度下半期の東証1部の出来高を1日あたり4、5億株と見込んで予算を立てており、薄商いが続くと、証券各社の苦境は深刻さを増しそうだ。

同日の1部の平均株価は先週末終値に比べて25.3円50銭安の2万4

【図4】



【図26】

計算結果

政治: $66.7 \times 3 + (-3.3) \times 3 + (-1.5) \times 3 + (-5.0) \times 4 + \dots = 165.7 + \dots$

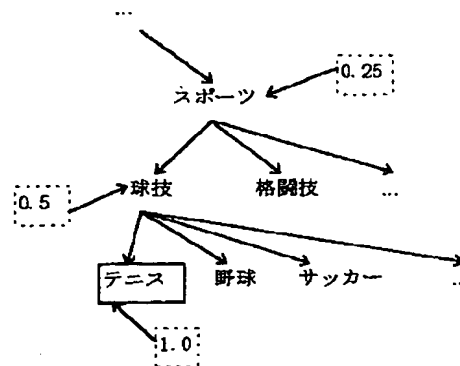
経済: $20.8 \times 3 + 9.0 \times 3 + 10.0 \times 3 + 9.0 \times 4 + \dots = 155.4 + \dots$

労働: $0.0 \times 3 + (-1.0) \times 3 + (-0.3) \times 3 + (-1.0) \times 4 + \dots = -8.1 + \dots$

スポーツ: $(-1.5) \times 3 + (-1.0) \times 3 + (-0.3) \times 3 + (-1.0) \times 4 + \dots = -12.9 + \dots$

国防: $1.5 \times 3 + 0.0 \times 3 + (-0.6) \times 3 + 18.0 \times 4 + \dots = 75.0 + \dots$

【図29】



【図33】

見出し文字列	品詞	題名	要目	...	合計	分類名
自然語	名詞	0	1	1	17	言語処理
辞書	名詞	1	2	1	21	言語処理
日本語	名詞	0	0	0	4	言語処理
...						
A/D変換	名詞	1	1	1	11	電気回路
アナログ	名詞	0	2	1	12	電気回路
...						

【図35】

(アメリカ, 政府, 先進, 主要, 国, ココム, ..., 意向)

アメリカ = (1, 1, 1, 1, 1, 1, ..., 0)

兵器 = (0, 0, 0, 0, 1, 1, ..., 1)

【図17】

(a)

首相(政治) = (5, 3, 5, 0, 0, ...)

首相(経済) = (0, 0, 3, 5, 8, ...)

内閣 (名詞) 国民 (名詞) 総裁 (名詞) 市場 (名詞) 経済 (名詞)

↓ 正規化

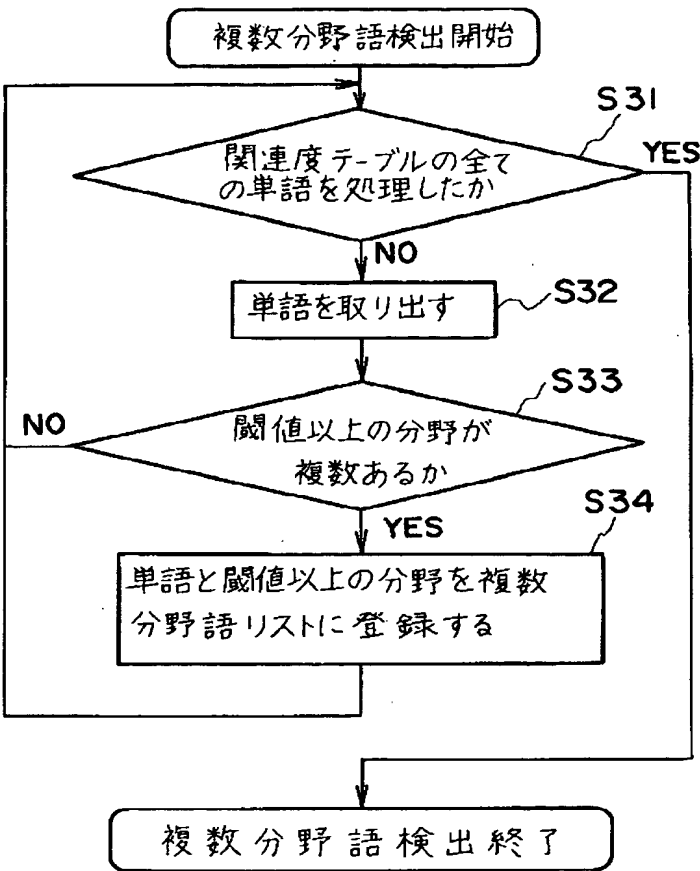
(b)

首相(政治) = (0.41, 0.25, 0.41, 0, 0, ...)

首相(経済) = (0, 0, 0.33, 0.55, 0.77, ...)

内閣 (名詞) 国民 (名詞) 総裁 (名詞) 市場 (名詞) 経済 (名詞)

【図 5】



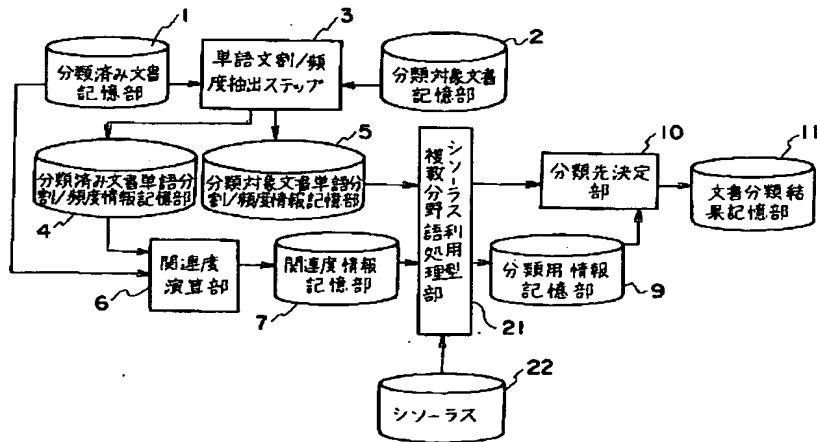
【図 3 6】

(アメリカ, 政府, 先進, 主要, 国, ココム, ..., 意向)

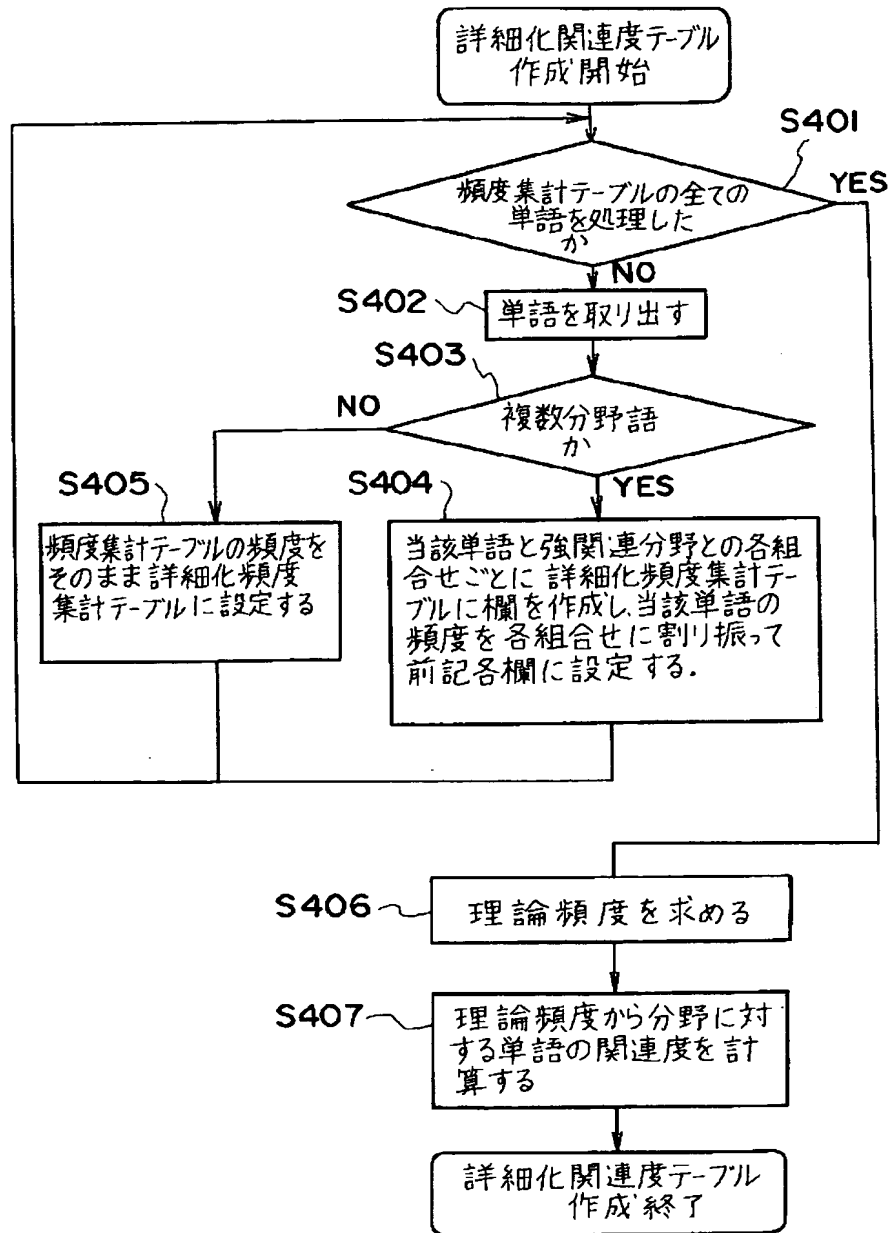
アメリカ	=	(1,	1,	1,	1,	1,	1, ...,	0)
兵器	=	(0,	0,	0,	0,	1,	1, ...,	1)

文書ベクトル = (1, 1, 1, 1, 2, 2, ..., 1)

【図 2 7】



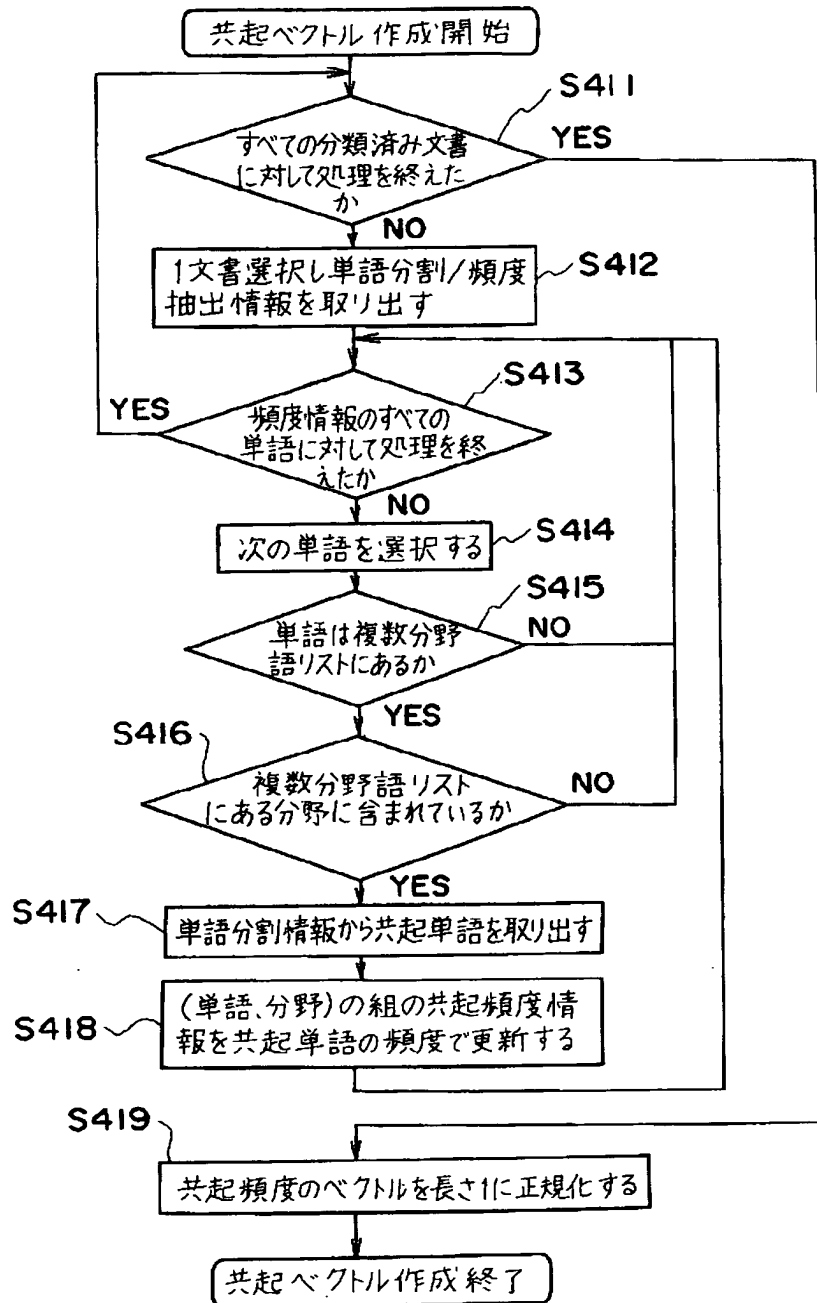
【図 6】



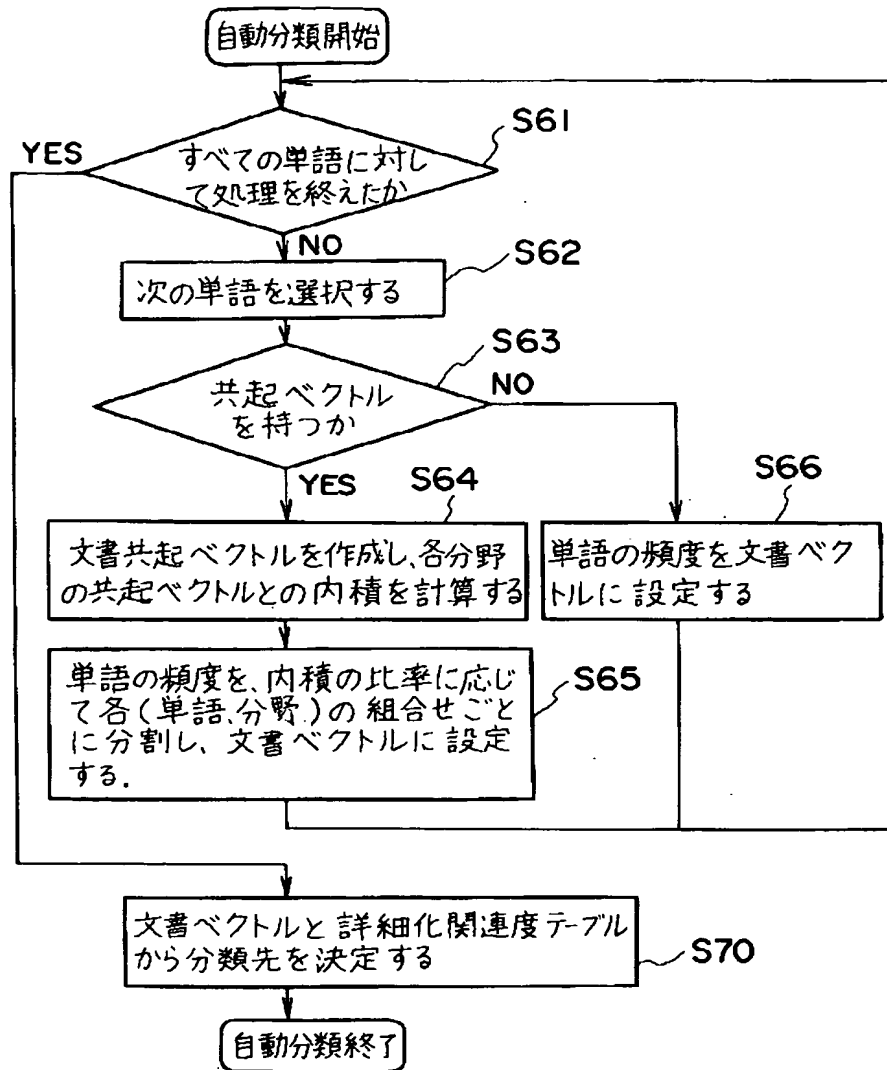
【図 3 8】

名詞	一般名詞意味属性	固有名詞意味属性
醤油 (一般名詞)	[調味料]	—
おかず (一般名詞)	[食料]	—
秋田 (固有名詞)	[行政区画]、[学校]	[県名]、[市名]、.....

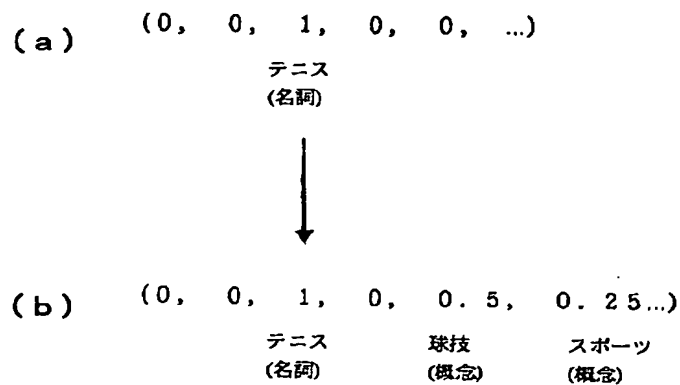
【図7】



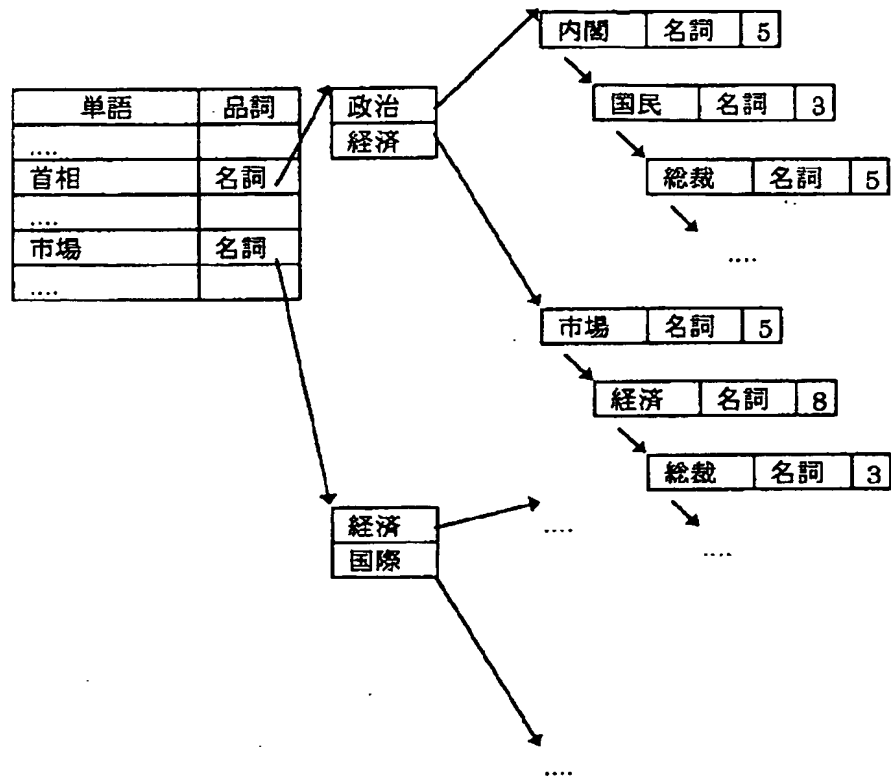
【図8】



【図30】



【図16】



【図31】

(a)

$$\begin{aligned} \text{首相(政治)} &= (0.41, 0.25, 0.41, 0, 0, \dots) \\ \text{首相(経済)} &= (0, 0, 0.33, 0.55, 0.77, \dots) \end{aligned}$$

内閣	国民	総裁	市場	経済
(名詞)	(名詞)	(名詞)	(名詞)	(名詞)

↓

(b)

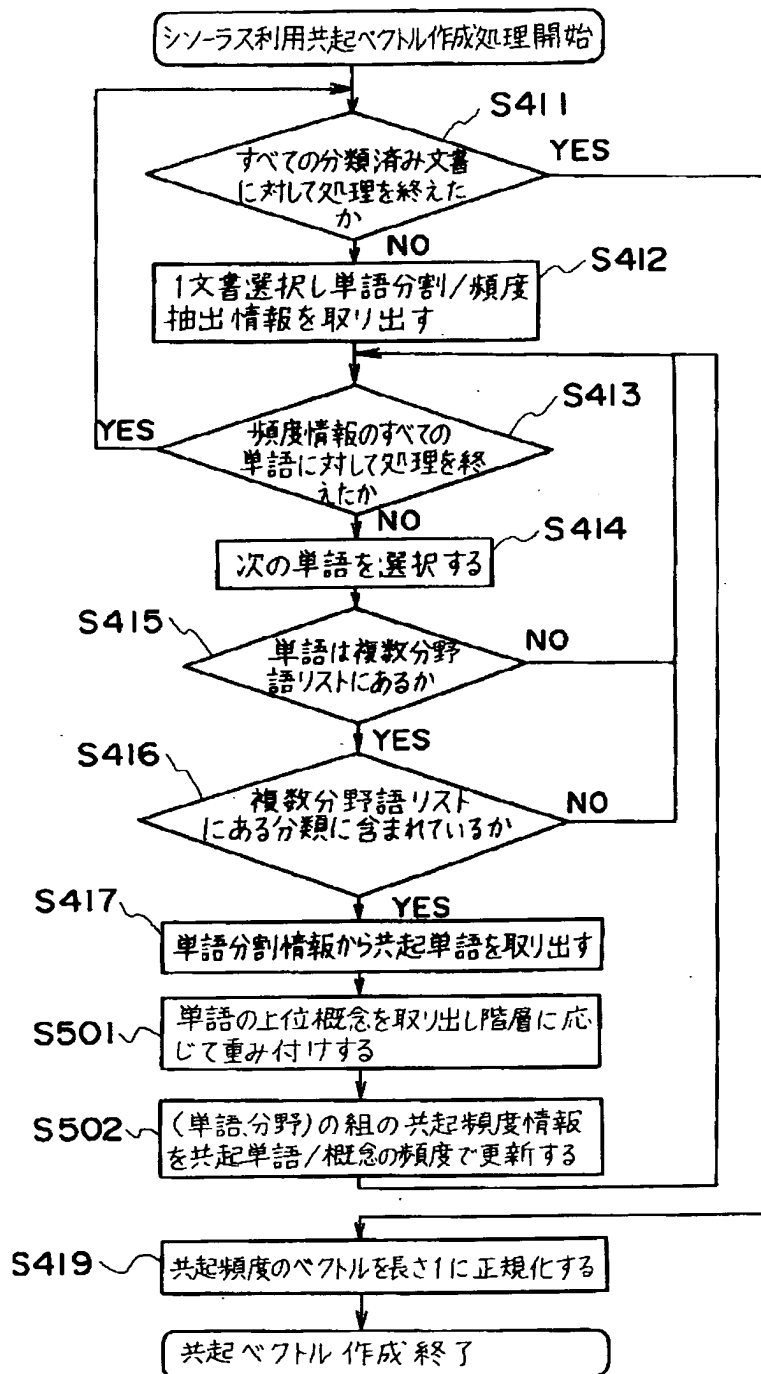
$$\begin{aligned} \text{首相(政治)} &= (0.41, 0.25, 0.20, 0, 0, \dots) \\ \text{首相(経済)} &= (0, 0, 0.16, 0.55, 0.77, \dots) \end{aligned}$$

総裁(名詞)

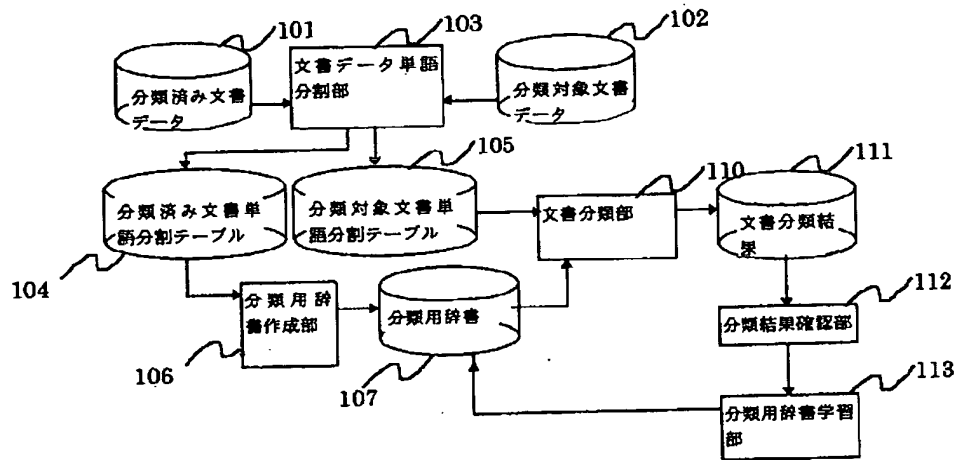
【図39】

意味属性	分野				
	政治	経済	...	家庭	運輸通信
[人工物]	7	10		12	12
[資材]	0	7		1	9
.....					
[食料]	1	1		9	1
[調味料]	0	0		7	1

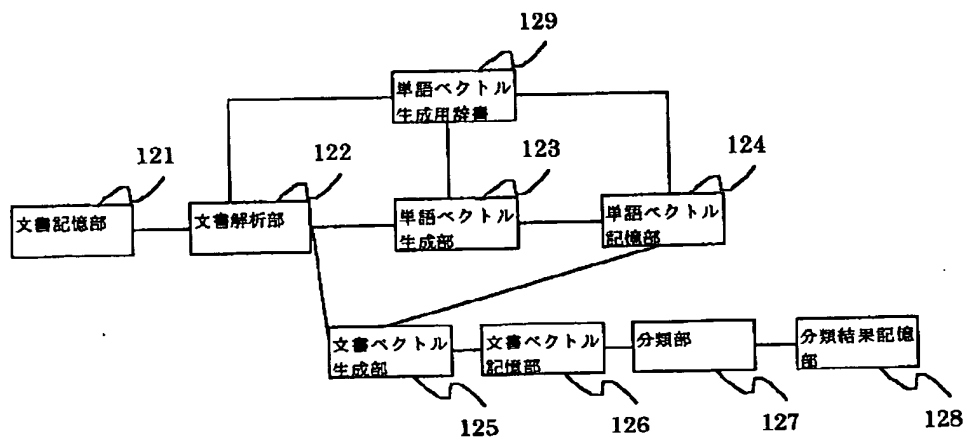
【図28】



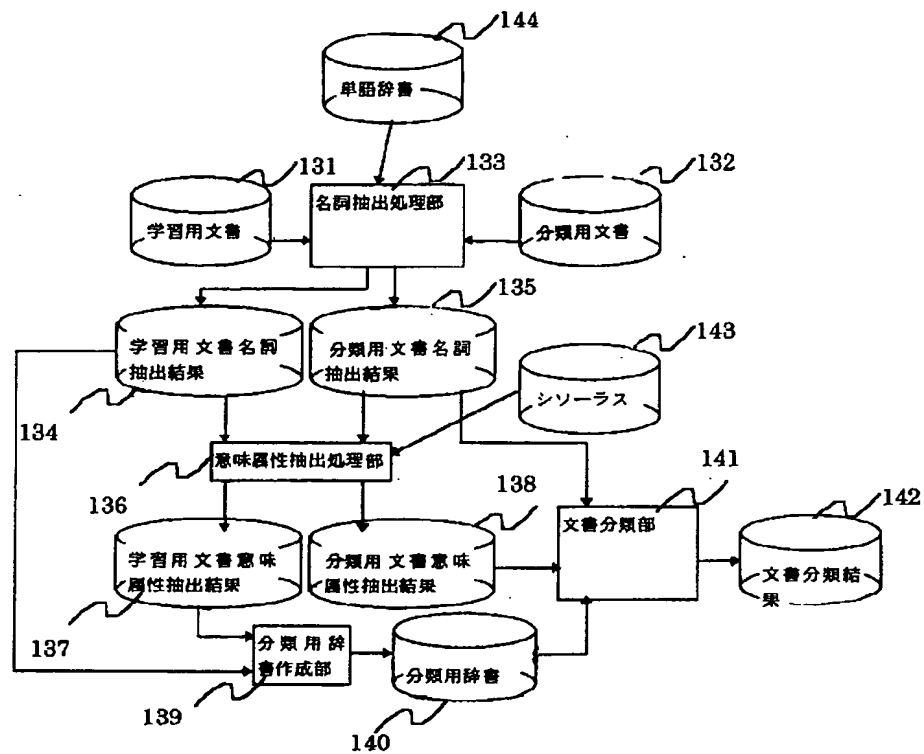
【図32】



【図34】



【図37】



フロントページの続き

(72)発明者 高山 泰博
 東京都千代田区丸の内二丁目2番3号 三
 菱電機株式会社内